CESifo DICE REPORT

Journal for Institutional Comparisons

VOLUME 11, NO. 1 SPRING 2013

Forum

MANAGING THE HEALTHCARE SYSTEM
Stefan Felder
Friedrich Breyer
Carol Propper
Jan Brosse and Mathias Kifmann
Peter Zweifel
Hans Maarse, Patrick Jeurissen and Dirk Ruwaard
Amanda Goodall

Research Reports

THE NETWORK EFFECT IN INTERNATIONAL MIGRATION
Michel Beine

FINANCING THE GERMAN ECONOMY
Christa Hainz and Manuel Wiegand

Reform Model

INCENTIVE SCHEMES FOR LOCAL GOVERNMENT
Ben Lockwood and Francesco Porcelli

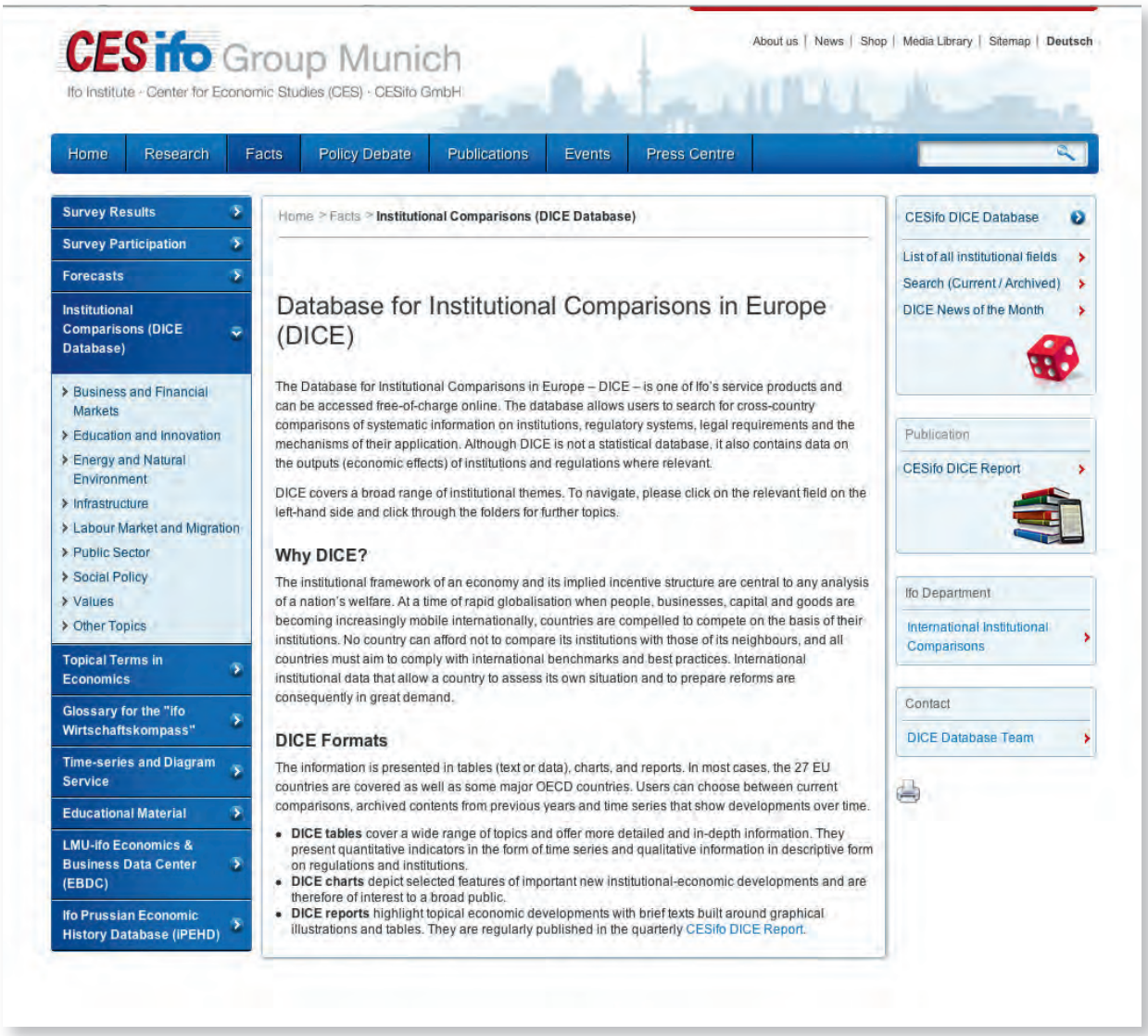
Database

THE BISMARCKIAN FACTOR
ULCS IN THE EUROZONE
GENDER WAGE GAP
ASYLUM SEEKERS

News

NEW AT DICE DATABASE, CONFERENCES, BOOKS

CESifo The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute, Munich



CESifo DICE Report
ISSN 1612-0663 (print version)
ISSN 1613-6373 (electronic version)
A quarterly journal for institutional comparisons
Publisher and distributor: Ifo Institute
Poschingerstr. 5, D-81679 Munich, Germany
Telephone ++49 89 9224-0, Telefax ++49 89 9224-1462, e-mail ifo@ifo.de
Annual subscription rate: €50.00
Editors: Marcus Drometer, Silke Friedrich
Editor of this issue: Marcus Drometer (drometer@ifo.de)
Copy editing: Lisa Giani Contini, Sabine Rumscheidt, Andrea Hohenadl

DICE Database: www.cesifo-group.org/DICE

CESifo DICE Report

Volume 11, Number 1 Spring 2013

Forum

Managing the Healthcare System

The Impact of Demographic Change on Healthcare Expenditure
Stefan Felder 3

Implicit Versus Explicit Rationing of Health Services
Friedrich Breyer 7

Competition, Incentives and the English NHS
Carol Propper 16

Competition in Health Insurance and Premium Regulation
Jan Brosse and Mathias Kifmann 21

Managed Care: Prescription for Failure? Lessons from Switzerland
Peter Zweifel 27

Concerns over the Financial Sustainability of the Dutch Healthcare System
Hans Maarse, Patrick Jeurissen and Dirk Ruwaard 32

Should Doctors Run Hospitals?
Amanda Goodall 37

Research Reports

The Network Effect in International Migration
Michel Beine 41

Financing of the German Economy During the Financial Crisis
Christa Hainz and Manuel Wiegand 48

Reform Model

Incentive Schemes for Local Government
Ben Lockwood and Francesco Porcelli 55

Database

The Bismarckian Factor 64
Unit Labour Costs in the Eurozone 67
The Gender Wage Gap in OECD Countries 69
Inflows of Asylum Seekers to OECD Countries 72

News

New at DICE Database, Conferences, Books 74

Managing the Healthcare System

The Impact of Demographic Change on Healthcare Expenditure[1]

Stefan Felder*

Introduction

Men and women in the developed world have experienced a significant increase in life expectancy over the last 50 years. At the same time per capita healthcare expenditure has increased dramatically. This joint trend led the OECD (1988) and others (e.g. Mendelson and Schwartz 1993) to blame population ageing for the increase in healthcare expenditure. Since healthcare expenditure is a rising function of age – from the age of 60 onwards it takes the form of an almost exponentially rising curve – part of the increase in healthcare expenditure may, in fact, be due to population ageing. Nonetheless, Zweifel, Felder and Meier (1999) found that age has no effect on healthcare expenditure after controlling for proximity to death. More recently, Shang and Goldman (2007) find that age has little additional predictive power on healthcare expenditure after controlling for remaining life expectancy and even the predictive power of life expectancy declines when health status controls are included in the analysis. This result is consistent with previous findings suggesting that the expected cumulative health expenditure for healthier elderly individuals, despite their greater longevity, is similar to that for less healthy persons (Lubitz, Beebe and Baker 1995).

In the next section (*The red herring hypothesis*) this article surveys the growing body of literature on the relationship between population ageing and healthcare expenditure, in particular on the red herring hypothesis (Zweifel et al. 1999), claiming that population ageing is even neutral with respect to the increase in healthcare expenditure. The third section (*Forecasting future health expenditures*) deals with predictions of future healthcare expenditure. The fourth section (*The role of increasing life expectancy*) emphasises the dynamic effects of increased life expectancy on healthcare expenditure, and the fifth section (*Ageing and health technologies*) addresses the relationship between population ageing and health technology. The final section offers some concluding remarks.

The red herring hypothesis

As soon as it became apparent that the OECD countries would experience severe population ageing, researchers and policy analysts advising politicians started warning of the threat of exploding expenditure for healthcare, because in a cross-section, higher age is associated with greater healthcare utilisation. Victor Fuchs was the first to observe that "healthcare spending among the elderly is not so much a function of time since birth as it is a function of time to death. The principal reason why expenditure rises with age in a cross-section (among persons aged 65 and over) is that the proportion of persons near death increases with age" (1984, pp. 151f.). But it took one and a half decades for this relationship to be explored more thoroughly using modern econometric techniques. In their pioneering study, Zweifel et al. (1999) analysed the expenditure of roughly 1,000 persons who had died in Switzerland in the period 1983 to 1992 and found that among those who died beyond age 65, healthcare expenditure in the last eight quarters of life did not depend significantly on calendar age, whereas it increased significantly with proximity to death. The authors also failed to find an age effect in years five to two before death and thus concluded: "Exclusive emphasis on population ageing as a cause of growth in per capita healthcare expenditure runs the risk of creating a red herring by distracting from the choices that ought to be made …" (p. 494).

The 'red herring hypothesis' was born. It was in perfect agreement with the compression-of-morbidity thesis by Fries (1980), which stated that the onset of disability is postponed and the time span of severe

* University of Basel and CINCH Essen.
[1] The article borrows from a recent survey by Breyer, Felder and Costa I Font (2011).

illness leading to death shrinks when life expectancy increases. While the Zweifel et al. (1999) study suffered from the weakness of concentrating on patients in their last two years of life, subsequent studies by several authors mainly confirmed the red-herring hypothesis. Felder, Meier and Schmitt (2000) analysed a subsample of the data set used by Zweifel et al. (1999) and demonstrated that for persons over 65 years of age, holding time to death constant, healthcare expenditure even decreased with age. Seshamani and Gray (2004a) showed that hospital costs in Britain start rising as early as 15 years before death, whereas the relationship between age and hospital costs is inversely U-shaped and peaks at age 80. Similarly, Seshamani and Gray (2004b) found that age has a small positive effect on hospital costs. O'Neill et al. (2000) found no age effect on the general practitioners' cost associated with nursing home patients when controlling for proximity to death.

Salas and Raftery (2001) argued hat proximity to death may be endogenous if healthcare interventions have a positive effect on the patient's health. Felder, Werblow and Zweifel (2010) addressed the endogeneity issue in an extended empirical study and showed that this does not change the main result: proximity to death and not calendar age is the crucial determinant of healthcare expenditure.

A further refinement of the analysis of age and healthcare expenditure was achieved in Werblow, Felder and Zweifel (2007). They decomposed expenditure into several components and found that the age pattern of expendiure not only differed considerably between survivors and decedents, but even more strongly between users and non-users of long-term care: while the age profile for deceased non long-term-care users was monotonically declining, surviving non-users had a hump-shaped profile peaking at age 80. On the other hand, users of long-term care had an increasing age profile even for acute healthcare expenditure, which is more pronounced for survivors than for decedents. These findings confer with those of Spillman and Lubitz (2000) who analysed the healthcare expenditure of the US Medicare population, i.e. individuals aged 65+. They report a convex (from below) age profile for both nursing home care and (less accentuated) for home care. By contrast, services covered by Medicare and prescription drugs exhibit a decreasing age profile. This implies a continuing shift from acute to long-term care late in life. Spillman and Lubitz conclude that population ageing will be an important driver of demand for long-term care, leaving the acute sector unaffected.

Forecasting future health expenditure

While the studies summarised above all try to explain the relationship between age and healthcare expenditure in past data, it may be argued that the true purpose of these exercises is to derive more solid predictions of the future development of healthcare expenditure. Indeed, it was shown in several studies that, taking time to death into account, expenditure forecasts become less dramatic. Stearns and Norton (2004) compared predictions of Medicare expenditure for the year 2020 on the basis of observed expenditure data from the period 1992 to 1998, which were inferred from different regression models. They found that neglecting time to death in the regression model leads to an overestimation of the expenditure increase by 15 percent. Polder, Barendregt and van Oers (2006) for the Netherlands found that including time to death led to a ten percent reduction in the growth rate of future health expenditure compared to conventional projection methods.

Breyer and Felder (2006) applied the estimated regression coefficients derived by Zweifel, Felder and Werblow (2004) to the projections of the age structure and mortality rates for the German population between 2002 and 2050 as published by the Federal Statistical Office. They found that compared to a 'naïve' projection, which uses the unadjusted age-expenditure profile, distinguishing explicitly between survivors' and decedents' healthcare expenditure dampens the projected increase up until 2050 by roughly 20 percent. Adding a 'compression-of-morbidity' assumption – stating that if life expectancy increases between 2002 and 2050 by x years, then, for example, a 65-year-old person in 2050 will be as healthy as a 65-minus-x-year-old in 2002 – lowers the expenditure projection by another 20 percent. The surprising result of this exercise is that, even accepting the 'red-herring' assumptions, there will still be a sizeable demographic effect on healthcare expenditure. This result was confirmed by Steinmann, Telser and Zweifel (2007), who calculated that taking the mortality effect into account lowers the forecast of the purely demographic effect on healthcare expenditure in Switzerland between 2000 and 2030 from an annual growth rate of 0.7 percent to 0.5 percent. The analysis nevertheless agrees that popula-

tion ageing has a positive effect on heath care expenditure.

The role of increasing life expectancy

An important weakness of almost all studies in the literature is their reliance on cross-section expenditure data. Therefore, in drawing inferences from these studies for the development of healthcare expenditure over time, proponents of the 'red-herring' hypothesis are subject to the same error of which they accuse their opponents (i.e. those who believe that ageing increases health spending because per-capita expenditure increases with age). In particular, they overlook the fact that increasing longevity not only means that 30 years from now the average age at death will be higher, but also that people at a certain age (say, 75) will on average have more years to live than current 75-year olds. As a consequence, future physicians will look at 75-year old patients with different eyes than those of present physicians, because the notion of a 'normal life-span' will have shifted upwards. This effect is consistent with the ethical justification of age-based rationing of healthcare services (Callahan 1987; Daniels 1985), and with the corresponding empirical literature, which shows that some physicians do indeed use age as a criterion in allocating scarce healthcare resources (for an overview see Strech et al. 2008).

Thus, to address the crucial question of how healthcare expenditure will react to population ageing, (i.e. an increase in life expectancy?), an econometric estimation of the determinants of expenditure has to be modified in two directions: firstly, by looking at panel rather than cross-section data, and secondly, by including a direct measure of remaining life expectancy as a regressor. Of course, this cannot be done with individual data, but requires an estimation with population group averages as units of observation. This approach has only been followed by two studies. The first one is Zweifel, Steinmann and Eugster (2005), which addresses the 'Sisyphus Syndrome' in healthcare, i.e. the mutual reinforcement of population ageing and public spending on healthcare of the elderly, by looking at a panel of OECD countries for the period 1970 to 2000. Remaining life expectancy weighted with the share of the population older than 65 turns out to be a significant and positive determinant of health expenditure as a share of GDP. This confirms the hypothesis that population ageing increases healthcare expenditure. Only the interpretation differs from the naïve one discussed above: it is not medical need, but rather political weight that explains why an older population demands a higher public spending on healthcare.

In a recent unpublished paper, Breyer, Lorenz and Niebel (2012) used data for a pseudo-panel of all German sickness fund members (grouped by age and gender) over the period 1997–2008. In a fixed-effects regression, they found that age, mortality rate and the remaining life expenctancy of persons over 60 have a positive impact on per capita healthcare expenditure. They then simulated future healthcare expenditure in Germany on the basis of an official population forecast including life expectancy and discovered that demographic change itself is associated with an annual growth rate of roughly 0.5 percent.

Ageing and health technologies

An important question in this context is whether medical progress predominantly benefits the aged. If this is the case, then the findings of the previous sections (that population ageing affects health care expenditure only weakly) must be regarded with great caution. One popular method to test this proposition is to look at whether age-expenditure profiles become steeper over time. A number of papers have addressed this question, but the answers are diverse and therefore inconclusive.

Buchner and Wasem (2006) analysed data from the largest private health insurer in Germany for the period 1979 to 1996 and defined three different indicators for a 'steepening' of the age-expenditure profile over time. In particular the increase in per capita healthcare expenditure of the 'old', using 65 as cut-off age, was significantly larger than the corresponding figure for the 'young'. Felder and Werblow (2008) challenged this result by looking at average expenditure data in the Swiss cantons over the period 1997 to 2006. In a panel regression with population averages as units of observation, the interaction effect of time and age group dummies was not consistently increasing in age. However, for all age groups between 65 and 90, this interaction effect was positive and significant at the 10 percent level. Thus even for Switzerland it cannot be ruled out that the increase in healthcare has recently been particularly large in those age groups that will rapidly increase in size over the next decades.

Conclusion

Population ageing is often blamed for the steady increase observed in healthcare expenditure in the Western world. Robert Evans (1985) suggested that the fixation on ageing provides an "illusion of necessity". By making it seem as though healthcare expenditure is inevitable in higher age, attention is diverted from the real causes of growth of the healthcare sector. These are technical progress in medicine, the secular increase in income, and wrong incentives for the providers and consumers of healthcare caused by government regulation and extensive social health insurance coverage. Rephrasing Evans, Zweifel et al. (1999) stated that blaming population ageing serves as a red herring, distracting from choices that ought to be made to curb steadily rising healthcare expenditure in the Western world.

Overall, empirical studies suggest that the impact of a longer life on future healthcare expenditure will be quite moderate because of the high costs of dying and the compression of mortality and morbidity in old age. If proximity to death, and not age per se, determines the bulk of expenditure, a shift in the mortality risk to higher ages will not significantly affect lifetime healthcare expenditure, as death occurs only once in every life. An exception to this rule is long-term care. As ever more people reach a very high age (beyond 85 or 90), the percentage needing long-term care in their last years of life increases.

A calculation of the demographic effect on healthcare expenditure in Germany up until 2050 that explicitly accounts for costs in the last years of life leads to a significantly lower demographic impact on per capita expenditure than a calculation based on crude age-specific health expenditure. The pure age-effect of population ageing on the annual growth rate of per capita healthcare expenditure does not exceed 0.5 percentage points, i.e. is much lower than the observed annual real growth rate of around two percent in the OECD.

References

Breyer, F. and S. Felder (2006), "Life Expectancy and Health Care Expenditures: A new Calculation for Germany Using the Costs of Dying", *Health Policy* 75 (2), 178–86.

Breyer, F., S. Felder and J. Costa I Font (2011), "Ageing, Health, and Health Care", *Oxford Review of Economic Policy* 26, 674–90.

Breyer, F., N. Lorenz and T. Niebel (2012), "Population Ageing and Health Care Expenditures: Is there a Eubie Blake Effect?", *DIW Working Paper* no. 1226, Berlin.

Buchner, F. and J. Wasem (2006), ""Steeping" of Health Expenditure Profiles", *The Geneva Papers on Risk and Insurance: Issues and Practice* 31 (4), 581–99.

Felder, S. and A. Werblow (2008), "Do the Age Profiles of Health Care Expenditure Really Steepen over Time? New Evidence from Swiss Cantons", *The Geneva Papers on Risk and Insurance: Issues and Practice. Special Issue on Health Insurance* 33 (4), 710–27.

Felder, S., A. Werblow and P. Zweifel (2010), "Do Red Herrings Swim in Circles? Controlling for the Endogeneity of Time to Death" *Journal of Health Economics* 29 (2), 205–12.

Felder, S., M. Meier and H. Schmitt (2000), "Health Care Expenditure in the Last Months of Life", *Journal of Health Economics* 19 (5), 679–95.

Fries, J. F. (1980), "Ageing, Natural Death, and the Compression of Morbidity", *New England Journal of Medicine* 303 (3), 130–36.

Fuchs, V. R. (1984), "Though Much is Taken: Reflections on Aging, Health and Medical Care", *Milbank Memorial Fund Quarterly/Health and Society* 62 (2), 143–66.

Lubitz, J., J. Beebe and C. Baker (1995),"Longevity and Medicare Expenditure", *New England Journal of Medicine* 332 (15), 999–1003.

Mendelson, D. N. and W. B. Schwartz (1993), "The Effects of Aging and Population Growth on Health Care Costs", *Health Affairs* 12 (1), 119–25.

OECD (1988), *Ageing Population: The Social Policy Implications*, OECD Publishing, Paris.

O'Neill, C., L. Groom, A. J. Avery, D. Boot and K. Thornhill (2000), "Age and Proximity to Death as Predictors of GP Care Costs: Results from a Study of Nursing Home Patients", *Health Economics* 9 (8), 733–38.

Polder, J. J., J. J. Barendregt and H. van Oers (2006), "Health Care Costs in the Last Year of Life – The Dutch Experience", *Social Science and Medicine* 63 (7), 1720–31.

Salas, C. and J.P. Raftery (2001), "Econometric Issues in Testing the Age Neutrality of Health Care Expenditure", *Health Economics,* 10 (7), 669–671.

Seshamani, M. and A. Gray (2004a), "Ageing and Health Care Expenditure: The Red Herring Argument Revisited", *Health Economics* 13 (4), 303–14.

Seshamani, M. and A. Gray (2004b), "A Longitudinal Study of the Effects of Age and Time to Death on Hospital Costs", *Journal of Health Economics* 23 (2), 217–35.

Shang, B. and D. Goldman (2007), "Does Age or Life Expectancy Better Predict Health Care Expenditures?" *Health Economics* 17 (4), 487–501.

Spillman, B. C. and J. Lubitz (2000), "The Effect of Longevity on Spending for Acute and Long-term Care", *New England Journal of Medicine* 342 (19), 1409–15.

Stearns, S. C. and E. C. Norton (2004), "Time to Include Time to Death? The Future of Health Care Expenditure Predictions", *Health Economics* 13 (4), 315–27.

Steinmann, L., H. Telser and P. Zweifel (2007), "Aging and Future Health Care Expenditure: A Consistent Approach", *Forum for Health Economics & Policy* 10 (2), 1–30.

Werblow, A., S. Felder and P. Zweifel (2007), "Population Ageing and Health Care Expenditure: A School of „Red Herrings"?", *Health Economics* 16 (10), 1109–26.

Zweifel P., S. Felder and M. Meier (1999), "Ageing of Population and Health Care Expenditure: A Red Herring?", *Health Economics* 8 (6), 485–96.

Zweifel, P., S. Felder and A. Werblow (2004), "Population Ageing and Health Care Expenditure: New Evidence on the "Red Herrings"", *The Geneva Papers on Risk and Insurance* 29 (4), 652–66.

Zweifel, P., L. Steinmann and P. Eugster (2005), "The Sisyphus Syndrome in Health Revisited", *International Journal of Health Care Economics and Financing* 5 (2), 127–45.

Implicit Versus Explicit Rationing of Health Services[1]

Friedrich Breyer*

"At least as long I am Minister of Health, I shall never lead a debate on rationing or prioritization, for ethical reasons" (Philipp Rösler 2010).

Introduction[2]

In many developed countries, the concept of rationing healthcare services is treated as a taboo in the political debate. If someone argues in favor of certain types of explicit rationing, s/he immediately encounters fierce reactions by politicians and medical leaders and is sometimes even treated as if s/he had proposed euthanasia. The quotation above from the former German Health Minister Rösler, a Free Democrat, shows that this attitude is widespread in all political parties. Physician representatives like the late president of the German Medical Association, Jörg-Dietrich Hoppe, usually draw a line between the concepts of rationing (which they oppose) and prioritization (which they advocate). But even the latter concept is harshly rejected by office-holding politicians.

The purpose of the present paper is to contribute to a more sober and rational debate on this extremely emotional topic. To this end, the next section *(two definitions of rationing)* compares the two most popular definitions of the term rationing with respect to health services and contrasts them with the general concept of rationing in economics. The third section *(the euphemism of "prioritization")* shall analyse its relation to the concept of prioritization. The fourth section *(levels and types of rationing)* defines different levels and types of rationing, while the fifth section *(rationing in practice: a comparison of England/Wales and Germany)* uses these terms for a comparison of two real-world rationing schemes. The sixth section *(how to replace implicit with explicit rationing)* subsequently discusses options for the further development of explicit rationing, and the last section offers some conclusions.

Two definitions of rationing

Rationing as "withholding necessary services"

In the political sphere, healthcare rationing is commonly understood as "withholding necessary medical services".[3] This definition is potentially useful only if the concept of a "necessary medical service" is well-defined. Moreover, it is critical that the term "withholding" can be applied whenever a service delivered to an individual is not financed by a third party such as a sickness fund or the taxpayer.

When is a medical service necessary? The answer depends upon what the consequences would be if the patient does not get the service. Is it:

- an immediate danger to life,
- the risk of a severe and lasting health impairment, or
- any, even only temporary, deterioration of health?

Similarly, a health service cannot be called necessary if it is not even suitable for improving a patient's health, and even if this is the case, what is the minimum expected benefit to call the service "necessary": is it, for example, the gain of a few weeks life expectancy in a critical health state? Moreover, should costs be considered in the definition of what is "necessary"? Ubel (2000, 25) argues against mixing "necessity" with cost-effectiveness, but would he

[1] The author is grateful to Marlies Ahlert, Thorsten Kingreen and Hartmut Kliemt for valuable comments on an earlier version of this paper.
[2] A related paper in German appeared as Breyer (2012). Other papers on the same topic are Althammer (2008), Breyer and Schultheiss (2002) and Kliemt (1996), (2010).
* University of Konstanz and DIW Berlin.

[3] See, for example, Zentrale Ethikkommission der Deutschen Ärztekammer (2000, A-1019).

stick to this opinion if the costs of a life extension by one month were to be one million Euro? And what about 10 million or 100 million Euro? This shows that the concept of "necessary services" is so vague that it would not be wise to base the definition of rationing on it, but it would be better to replace it with a more meaningful term such as "useful services", as Buchanan does (1996, 335–36).[4]

The term "withholding" for "not giving free of charge" is equally problematic. Firstly, it contains an implicit value judgment because it suggests that the person from which something is "withheld" has a legitimate claim to the goods or service in question. Not only do value-laden words impede rational discussions, but in this case the reference to an (previously existing) claim is based on a misunderstanding because the very act of rationing can serve as a justification of legal claims to services; and thus the term should not presuppose the existence of those claims to begin with.[5] Consequently, Ubel (2000, 28) avoids this error when he defines "healthcare rationing" as "implicit or explicit mechanisms that allow people to go without beneficial services"

Rationing as "limited allocation"

The second error in equating rationing with withholding lies in the fact that it is not compatible with the textbook definition of the concept of rationing in Economics. There, "rationing" is defined either as synonymous with "allocation" or as a specific type of allocation. Some textbooks use the term rationing for any kind of determination of how scarce goods are distributed among competing uses or users. In this vein, Case and Fair (2008, Chapter 4) attribute a "rationing function" to the price, and Samuelson and Nordhaus (2001, 61) write: "… *competitively determined prices ration the limited supply of goods among those who demand them.*"

Summarizing this reasoning, it is useful to distinguish between a wide and a narrow sense of the word "rationing". In its wide sense, rationing coincides with "allocation" and refers to any method to determine who receives what quantity of a scarce good or service. These methods can be divided into those that make use of the price mechanism ("price rationing") and those that do not ("non-price rationing"), the latter being synonymous with rationing in its narrow sense. More specifically, this latter concept can be defined as the *allocation of limited amounts below market price*, which often means "free of charge". An allocation below market price implies that somebody else – the government or the community of insured people – bears the difference to the supply price. Rationing thus presupposes some kind of collective financing of the good in question.[6] This, in turn, precludes an unlimited allocation, in particular the provision of "optimal diagnosis and treatment" at the public's expense because, as Victor Fuchs (1984, 1572) states, "*No nation can provide 'presidential medicine' for all its citizens.*" The term "optimal treatment" refers to all services with a positive medical benefit, no matter what their costs are.

This implies that, in publicly financed health systems, the state must decide on the criteria by which the allocated quantities are limited. For healthcare services, common criteria for rationing are medical urgency, cost-effectiveness and sometimes waiting time. However, even if only part of all citizens have received positive allotments of a collectively financed resource, this does not necessarily imply that all others have to go without. On the contrary, it is conceivable that there are other ways in which citizens can procure the resource (at market price), either in a legal market for private treatment or by travelling abroad (see *Levels and types of rationing* for further details).

The euphemism of "prioritization"

As mentioned above, medical officials try to avoid the "R word", at least in public debates, and prefer to talk about prioritization. According to the Oxford Dictionary, the verb "to prioritize" has two meanings: 1) to designate or treat (something) as being very or most important, and 2) to determine the order for dealing with (a series of items or tasks) according to their relative importance. Prioritization 2 is somehow a prerequisite for prioritization 1: you need to have an order before you can privilege some

[4] "*Rationing – which means the withholding of care expected to be of net benefit – occurs throughout every healthcare system and is unavoidable*".

[5] This error in reasoning has already been criticized by Jeremy Bentham (1843, Article 2): "*But reasons for wishing there were such things as rights, are not rights; – a reason for wishing that a certain right were established, is not that right – want is not supply – hunger is not bread.*"

[6] At the moment of utilization, even private insurance companies allocate the good below market price. The insurance contract grants the right to participate in this rationing process.

item. Moreover, giving priority (higher rank) to something is equivalent to giving posteriority (lower rank) to all competitors; but nobody likes to talk about that because "prioritization" sounds better.

Prioritization as a prerequisite to rationing

If "rationing" is understood as the "limited allocation of health services", it opens up the question what rules the allocation process should follow. A plausible and transparent procedure for determining an allocation rule is to start with compiling a rank order of services (defined by illness type, patient group or treatment type) on the understanding that this rank order will be followed in the allocation process from the top down until the capacity is fully exhausted, or the available funds are fully spent. In this interpretation, prioritization is an important first step towards a (rational) rationing process.

The most famous example of such a combination of prioritization and rationing is the Medicaid program of the state of Oregon in the US in the early 1990s (see Garland 1992). In the Oregon Basic Health Services Act of 1989 a rank order of 709 disease-treatment pairs for Medicaid beneficiaries according to urgency was compiled. In 1991 the funds that were allocated by the state to the Medicaid program were sufficient to finance only 587 of these 709 services.

Prioritization as an alternative to rationing

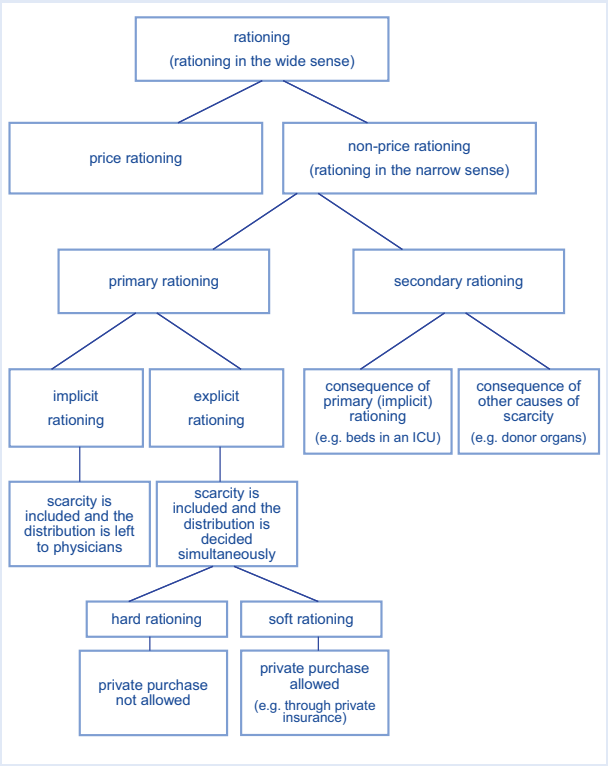
In contrast to this interpretation, medical leaders often understand prioritization as a substitute for rationing, which they define as the withholding of services. As an example, the president of the German Medical Association, Frank Ulrich Montgomery (2011) referred to the swine flu pandemic of 2009. He first emphasized that everybody could have received the vaccination (no rationing), and subsequently explained that certain risk groups and groups that could have passed on the virus to others (e.g. medical personnel) were prioritized because the vaccine became available only gradually over time. In this example, prioritization applies only to a temporal sequence of service delivery so that eventually everybody would get treated at public expense. In the same interview, Montgomery justified the need to prioritize with the scarcity of funds in the healthcare budget and the necessity "to allocate the limited funds in a just way". But the latter case would imply that some patients near the bottom of the priority list would have gone without. This can be for two different reasons:

1. The patients would have been cured even without the service because the illness was only temporary. In this case, the cost-effectiveness of the treatment is questionable and it is debatable whether the treatment should have belonged to the benefit package of social health insurance to begin with.
2. The patients would have died from the disease: this implies that the treatment would have been necessary, and what Montgomery calls prioritization was, in fact, rationing, at least by his own use of the word.

Finally, there is the case whereby a rank order of urgency is compiled, but is not used as a basis for rationing decisions (because everybody gets the service anyway). In this case, prioritization is a useless

Figure 1

Types of rationing



Source: The author.

task which, if it requires any scarce resources, should remain undone.

Levels and types of rationing

Even if it is agreed that rationing is understood as limited allocation of collectively financed services below market price, it is useful to distinguish several levels and types of rationing (see Figure 1) under the headings primary – secondary, hard – soft and explicit – implicit.

Primary versus secondary rationing

Many authors use the term rationing exclusively to refer to the allocation of non-augmentable resources such as donor organs or beds in an intensive care unit. This level of rationing, which Calabresi and Bobbitt (1979) call "second-order tragic choices", and which we have therefore named "secondary rationing" (Breyer and Schultheiss 2002), is mainly characterized by questions of distributive justice: Shall the only available liver be given to the alcoholic or to the young woman who has fallen off a horse? Or shall the owner of a donor card be privileged in organ allocation to a person who has explicitly refused to donate his organs? Shall the last free bed in an ICU be given to the patient with the greatest risk of dying, to the one with the largest probability of success or to the one who has waited the longest time?

Choices like these will always be unavoidable, no matter how large capacities in healthcare grow. But for this very reason, they show that rationing cannot be equated with withholding: if there is only one donor heart available with two potential recipients and if it is given to one of them, who would claim that it is "withheld" from the other patient? Moreover, these decisions require value judgments and are least accessible to health economics reasoning.

In contrast, "primary rationing" (what Calabresi and Bobbitt call "first-order tragic choices") means that society deliberately limits the collectively financed resources for healthcare services because these services compete with other uses such as education, infrastructure or even private consumption. Unlike secondary rationing, which is one of the consequences of scarcity *within* the healthcare system, primary rationing is concerned with determining the level of scarcity of resources *for* the healthcare sector in response to the general scarcity of resources. These decisions are unavoidable as well, ever since medicine became so successful that it would, in principle, be possible to spend (nearly) all of GDP on *useful* health services. The question of what part of GDP to devote to publicly financed health services and what procedure to use to decide this matter is predominantly a question of efficiency and can therefore be analyzed in economic terms.

It must be emphasized that in a society which has neither a tax-financed national health service nor a mandatory social health insurance, primary rationing is not an issue because rationing presupposes the allocation of health services through some collectively financed institution.[7] In a purely privately financed health system, in which each consumer decides on his/her healthcare utilization – either directly or by signing an insurance contract – there is no point in a public discussion on the rationing of services (since rationing occurs only as individual self-rationing).

Hard versus soft rationing

Once a society has introduced a collectively financed healthcare system with (primary) rationing, two further principal decisions have to be taken. The first concerns the question of whether markets for those services shall be allowed that are not offered by the public system. If this is the case, we speak of "soft" rationing, otherwise of "hard" rationing.[8]

An example of hard rationing in practice is organ allocation, as laws everywhere prohibit markets for organs. Some authors argue in favor of extending this rule to all healthcare services in order to achieve equality of access to these services (see, for example, Krämer 1989, 87). In fact, many people share the judgment that it should be possible to buy a nicer, but not a longer life. However, it is questionable if this noble goal can ever be achieved in practice as there are major obstacles to it:

- the principle of a free society, which must accept that citizens have different desires and should be allowed to fulfill them as long as they bear the corresponding costs and do not harm others;

[7] In the United States, the rationing debate obtains its relevance through the tax-financed Medicare and Medicaid programs and through tax subsidization of health insurance premiums.
[8] Breyer and Kliemt (1994) introduced the terms "weak" and "strong rationing" for the same contents.

- the fact that markets exist for a variety of non-medical goods, which are highly relevant to a long and healthy life – sometimes even more so than medical services – such as healthy food, safer cars or healthy residential areas;
- the fact that even if the government was willing to ban markets for supplementary health services, in a world with open borders purchases abroad could not be prevented. Moreover, the ensuing "surgery tourism" would not only be inefficient due to unnecessary travel costs, but would primarily be used by the well-to-do, which is contrary to the goals of those who are in favor of hard rationing in the first place.

The result of soft rationing is some form of two-tier medicine, but one in which everybody has the right to choose the tier s/he wants to belong to.

Implicit versus explicit rationing

The second principal decision can be characterized as the choice between the following alternatives:

1. "Implicit" or "bedside rationing": here society only determines the share of GDP that is financed by taxes or mandatory contributions and devoted to the healthcare sector, but leaves it to physicians to allocate services to individual patients, particularly in the case of competing needs. Besides a global budget for the healthcare system as a whole, individual budgets for healthcare providers like hospitals are a typical instrument in this type of rationing.
2. "Explicit rationing": here society enacts precise and transparent rules that determine the circumstances under which certain persons can claim certain medical services. All services that are claimed must be financed so that, at least in the short run, total healthcare expenditure and hence tax rates cannot be fixed a priori.

Many people prefer implicit over explicit rationing because the former allows upholding the belief that death is always due to an unhappy fate, and never the result of specific rationing decisions, including one's own decision not to include a certain service in one's insurance contract (Hall 1994). Furthermore, it is argued that implicit rationing allows physicians to consider the specifics of each individual patient when taking their treatment decisions to a greater degree than rationing according to strict rules (see, for example, Mechanic 1992 and Hunter 1995).

This appeal to professional judgment is convincing if a number of conditions are fulfilled:

1. There is a consensus in society that a good criterion for the success of treatment is the expected benefit, measured, for example, in the quality adjusted life years (QALYs) gained.
2. All members of society have identical preferences with respect to length of life (in QALYs) and consumption.
3. The correlation between success of treatment and objectively measurable criteria such as chronological age is small.
4. Physicians dispose of a set of medical criteria (such as blood pressure, ECG), which, taken together, enable a fairly accurate forecast of the success of a treatment, whereas individual treatments cannot be operationalized well enough to base general allocation rules on them.

If conditions 1 and 2 are fulfilled, the appropriate criterion for including a service in the coverage of a collectively financed health insurance system is the cost per QALY ratio. If 3 and 4 are fulfilled as well, then the maximization of QALYs gained can be achieved by specifying a budget and letting physicians decide on the allocation of services among patients strictly according to medical criteria. As a result, the expected utility of the insured – as assessed behind the veil of ignorance – will be maximized.

It is obvious that some of these conditions are quite unrealistic. In particular, it is hard to dispute that people differ in their preferences for length of life versus standard of living. Furthermore, as physicians can be influenced in their decisions, there is the danger that better educated and more eloquent patients are favored in rationing decisions taken at the bedside.

An additional weakness lies in the way in which implicit rationing is often achieved in practice, namely by limiting medical capacity. Although this practice has the advantage that physicians do not have to deny individuals a treatment despite the availability of sufficient resources to perform the treatment (see, for example, Krämer 1993, 55 ff.), there is a significant disadvantage attached to it. Most of the bigger countries like Great Britain or Germany are divided into regional units, which act as service areas for medical capacity, and it is practically impossible to align capacity perfectly with demand for services in every region. Falling short of this target, however,

implies a problem of inequality; since equal demand will meet unequal intensity of treatment in different regions, so that the principle of horizontal equity is jeopardized.

A further objection against implicit rationing is that it is hard to see why the task of distributing survival chances should be delegated to physicians for the sole reason that they possess the technical knowledge of what specific services are necessary to achieve this survival. In particular, their superior technical competence does not at all give physicians a superior moral competence for placing relative values on human lives (Kliemt 1993, 266). Interestingly, this argument is often made by physicians themselves (see, for example, Loewy 1991).

Moreover, the potential advantages of *soft* rationing can only be achieved if it is also explicit, i.e. if it is clear to every citizen which services are covered by Social Health Insurance or a National Health Services and which are not, so that a supplementary private insurance contract could cover the latter.

Finally, it is a consequence of the rule of law that whenever the government uses coercion to influence citizens' behavior, it is obliged to define the rights and duties of those citizens clearly, so that they can be reviewed by the courts of justice. This principle is violated in the case of implicit rationing whereby insurance coverage does not guarantee a claim to specific medical services in every single case.

Rationing in practice: a comparison of England/ Wales and Germany

Healthcare rationing in the NHS of England and Wales

The National Health Service (NHS) has always been the prototype of healthcare rationing. Healthcare provision through the NHS is completely tax-financed and the performance rates of certain medical procedures such as X-rays or renal dialysis per capita used to be only a fraction of the rates in the USA (Aaron and Schwartz 1984, 33, 73). Moreover, explicit rationing criteria such as age seem to have played a role for a long time,[9] and for elective procedures such as hip replacement waiting lines were used (ibid., 58–61), which are also explicit in the sense that the patient knows why s/he is not getting the service immediately and can, in principle, purchase it in a market for private healthcare services.

As an additional explicit rationing criterion, cost effectiveness started to play an increasing role under the Labour Government of 1997–2009. In 1999, the "National Institute of Clinical Excellence (NICE)"[10] was created "*to promote clinical and cost-effectiveness by producing clinical guidelines and audits, for dissemination throughout the NHS*" (Nelson 2011, 210). The main purpose of NICE is to appraise the cost-effectiveness of new drugs or medical procedures on the basis of scientific evidence and make recommendations to regional health authorities (called "Primary Care Trusts", PCTs).

The criterion used to arrive at a verdict is the "incremental cost effectiveness ratio" (ICER), which measures the additional costs and benefits, as compared to the best already available drug or procedure. Benefits are usually measured in "quality adjusted life years" (QALYs) gained, and a drug is approved if its ICER lies below a cost-per-QALY threshold. More precisely, PCTs are recommended to finance the drug if the ICER lies below 20,000 GBP, to give additional reasons if it lies between 20,000 and 30,000 GBP and to refuse financing if it exceeds 30,000 pounds (Walker, Palmer and Sculpher 2007, 56). So, at least as far as the use of pharmaceuticals is concerned, rationing is explicit in two respects: firstly, PCTs clearly state which drugs they do or do not finance; and secondly, the criterion used to justify the decision is also transparent.

The coalition government in power since 2009, however, announced that it would withdraw NICE's power to decide that drugs should not be provided based on cost-effectiveness determinations and introduce a new regime of negotiated drug pricing instead. Nelson (2011, 211–12) sees this as a clear indication of a transition from explicit rationing with transparent criteria to implicit rationing.

Healthcare rationing in German Social Health Insurance

In the German Social Health Insurance (SHI), explicit rationing is hardly ever used in the funding

[9] Aaron and Schwartz (1984, 34-37) report that this was true at least in allocating places for renal dialysis although physicians tried to conceal the fact that age as such was decisive.

[10] Later it was renamed as the "National Institute for Health and Clinical Excellence" without changing its acronym.

decision for new drugs and procedures. According to § 12 SGB V, services must be "sufficient, appropriate and economical, and they must not exceed the necessary quantity". If there is no appropriate alternative to a drug, it is automatically included in the benefit package of SHI. In 2004, an element of explicit rationing was introduced into the drug approval rules. An institute was founded with a similar design to that of NICE, the "Institute for Quality and Efficiency in Healthcare" (IQWiG), and it was commissioned to develop procedures for health technology assessment. In the first draft of these procedures, which were issued in early 2008, it was proposed to introduce a price ceiling for new drugs that should be defined by the relevant part of the "efficiency frontier" of competing drugs already in the respective market. In particular, the incremental cost-benefit ratio of the two best drugs in the market should be used to determine a price ceiling for the new drug. This would have been a clear case of explicit rationing because if the supplier of the new drug had refused to offer the drug at this price, it would not have been made available to members of SHI. However, a new law (AMNOG) in place since 2011 removed this possibility.

Decisions on the (non-)inclusion of new drugs or procedures in the benefit package of SHI are taken by the "Federal Joint Commission" (*Gemeinsamer Bundesausschuss, G-BA*), which comprises representatives of sickness funds and healthcare suppliers. In principle, this commission could reject a new drug if its extra benefit were to be deemed too small relative to its costs, compared to the next best alternative. In practice, this has never happened because the GB-A interprets the term "economical" in such a way that this requirement is always fulfilled if there is an additional benefit through the new drug, no matter how much extra it costs (Wasem 2012). If anything, the G-BA has in the past postponed the decision on the funding of a new drug, sometimes by several years, and has thus resorted to a kind of temporary explicit rationing (ibid.)

In the absence of explicit rationing devices, Germany uses a variety of regulations to contain healthcare expenditure such as a global budget for all ambulatory services, reference values for prescriptions and so-called efficiency checks, which force physicians to make decisions on the allocation of scarce resources (not least their own time). The criteria used to make such decisions were recently examined in surveys (see, for example, Schultheiss 2004, for a meta-analysis see Strech, Synofzik and Marckmann 2008). The authors show that it is not always medical criteria that determine physicians' decisions, but also contextual and individual factors like a patient's ability to articulate his/her wishes. A negative side effect is that rationing occurs not only implicitly, but is also concealed since the physician who must not lose the patient's trust will try to suggest that s/he has done everything to treat the patient in the optimal way.

A somewhat different approach was used by Thielscher, Schüttpelz and Schütte (2012) to quantify the extent of rationing related to patients suffering from one specific illness (schizophrenia). They determined the amount of time that a psychiatrist devoted to each patient in the year 2010, given the SHI reimbursement rates, and compared the result (10 minutes per month) with the time recommended by the respective clinical guideline (50 minutes per month). As the former number falls short of the latter one, the authors conclude that the services in question are rationed.

It is worth noting the ethically questionable fact that the limitations described in both studies, which are not caused by the objective unavailability of a well-defined resource (such as a transplant), are not practiced with respect to privately insured patients.[11] This means that there is two-tier medicine not only in the financing of, but also in the delivery of healthcare, and most citizens cannot even choose their affiliation to a specific tier.

How to replace implicit with explicit rationing

The considerations above suggest that it would be desirable to move towards explicit rationing, and to limit the extent of its implicit counterpart. This requires specifying the benefit package of SHI much more explicitly to create transparency for patients, healthcare providers and sickness funds. To be both practicable and acceptable to the public, the criteria for inclusion in the benefit package must not discriminate against well-defined patient groups, physicians must be willing to abide by the rules and finally – to create legal certainty – the criteria should be based on objective data and leave as little discretion as possible to the physicians who have to apply them.

[11] On the contrary, because of higher remuneration, private patients are often over-doctored.

Possible rationing criteria

In discussions on rationing, the following criteria play a major role:

- *Cost-effectiveness:* this criterion, which is the overriding one in defining the benefit package of the National Health Service in England and Wales, is attractive from the "behind the veil of ignorance" viewpoint because it maximizes expected quality-adjusted life expectancy from a given healthcare budget for the (still healthy) citizen. It is even favored by bio-ethicists as a result (see, for example, Marckmann and Siebert 2002). It might, however, discriminate against people with congenital diseases that are expensive to treat like haemophilia.
- *Patient age:* this criterion, which was allegedly used in the NHS in the 1960s and 70s, has the advantage of being operational and therefore facilitating supplementary insurance (Breyer and Schultheiss 2002). Physicians obviously accept it because they already apply it in situations of implicit rationing. In contrast, it seems to be a social taboo because many people think it is discriminating.[12]
- *Novelty:* as the increase in healthcare spending seems to be driven primarily by medical progress, an effective means of curbing this rise would be the delayed introduction of innovative pharmaceuticals and procedures (Häussler and Albrecht 2010). The disadvantage of this criterion is its weak ethical basis. Moreover, it is unclear whether it should also hold for cost-saving innovations. If not, there is little difference to the cost-effectiveness criterion.

Procedures of decision-making

Besides criticism of the prevailing implicit rationing as such, lawyers such as Kingreen (2011), also find fault with the fact that major decisions on distributing scarce resources are taken by a body such as the Federal Joint Commission, which is not legitimized by democratic procedures. In principle, the basic rules should be determined by parliament.

Of course, such decisions should follow a phase of open debate in public, i.e. in the media and in political parties, on strategies to cope with scarcity of resources in the publicly financed healthcare sector. In this debate, the preferences of the citizens regarding the trade-offs between length of life and consumption should be taken into account.[13] However, an indispensable prerequisite is the confession of politicians in all countries that rationing is unavoidable; and that no healthcare system, no matter how expensive it is, can guarantee all potentially beneficial services to all patients at the tax-payer's expense.

Concluding remarks

Everywhere in the world healthcare services are allocated in limited amounts, i.e. rationed. In modern welfare states, this allocation occurs independently of an individual's willingness or ability to pay, which means that there is rationing in the narrow sense of the word. Unfortunately, politicians (and even physician representatives) usually declare rationing as a taboo and thereby impede an open and public debate on the topic. Moreover, the euphemism of "prioritization" does not help to objectify the discussion, but rather tends to obfuscate it.

Instead of the prevailing implicit and often concealed rationing at the bedside, a free society under the rule of law needs explicit soft rationing provided by a well-specified Social Health Insurance benefit package. In the literature on this topic, several potential rationing criteria have been proposed. Societies, and eventually parliaments, should lead an open and honest debate of these criteria.

[12] This is a popular misunderstanding because age-based rationing actually achieves fundamental equality of treatment since age is not an invariant characteristic of a person like gender, but a series of states that every person passes through.
[13] A recent attempt to elicit people's willingness to pay for longer and healthier lives in nine European countries was made in the EuroVaQ project (Donaldson et al. 2010).

References

Aaron, H. J. and W. B. Schwartz (1984), *The Painful Prescription: Rationing Hospital Care*, Brookings Institution, Washington, D.C.

Althammer, J. (2008), "*Rationierung im Gesundheitswesen aus ökonomischer Sicht*", *Sozialer Fortschritt* 12, 289–94.

Bentham, J. (1843), "Anarchical Fallacies", in J. Bowring, ed., *The Works of Jeremy Bentham*, vol. 2, William Tait, Edingburgh, Article 2.

Breyer, F. (2012), "Implizite versus explizite Rationierung von Gesundheitsleistungen", *Bundesgesundheitsblatt* 55, 652–59.

Breyer, F. and H. Kliemt (1994), "Lebensverlängernde medizinische Leistungen als Clubgüter?", in K. Homann, ed., *Wirtschaftsethische Perspektiven I*, Duncker & Humblot, Berlin, 131–58.

Breyer, F. and C. Schultheiss (2002), "'Primary' Rationing of Health Services in Ageing Societies: A Normative Analysis", *International Journal of healthcare Finance and Economics* 2, 247–64.

Buchanan, A. (1997), "Health-Care Delivery and Resource Allocation", in R. M. Veatch, ed., *Medical Ethics*, Jones and Bartlett, New York, 321–61.

Calabresi, G. and P. Bobbitt (1978), *Tragic Choices,* Norton, New York.

Case, K. E. and R. C. Fair (2008), *Principles of Economics*, 8th ed., Pearson, Upper Saddle River, New Jersey.

Donaldson, C., R. Baker, H. Mason, M. Pennington, S. Bell, E. Lancsar … and P. Shackley (2010), *European Value of a Quality Adjusted Life Year*, final publishable Report of the EUROVAQ project, http://research.ncl.ac.uk/eurovaq/EuroVaQ_Final_Publishable_Report_and_Appendices.pdf.

Fuchs, V. R. (1984), "The "Rationing" of Medical Care", *New England Journal of Medicine* 311, 1572–73.

Garland, M. J. (1992), "Rationing in Public: Oregon's Priority-Setting Methodology", in M. A. Strosberg, J. M. Wiener and R. Baker, eds., *Rationing America's Medical Care: The Oregon Plan and Beyond,* Brookings Institution, Washington, D.C., 37–59.

Häussler, B. and M. Albrecht (2010), "Eine Versicherung für den Fortschritt", in B. Häussler, T. Isenberg, N. Klusen and A. Penk, eds., *Jahrbuch der medizinischen Innovationen*, Band 6: Innovation und Gerechtigkeit, Schattauer, Stuttgart, 81–86.

Hall, M. A. (1994), "The Problems with Rules-Based Rationing", *Journal of Medicine and Philosophy* 19, 315–32.

Hunter, D. J. (1995), "Rationing of Health Care: The Political Perspective", *British Medical Bulletin* 51 (4), 876–84.

Kingreen, T. (2011), "Knappheit und Verteilungsgerechtigkeit im Gesundheitswesen", in W. Höfling, ed., *Der Schutzauftrag des Rechts*, Veröffentlichungen der Vereinigung der Deutschen Staatsrechtslehrer, de Gruyter, Berlin/Boston, 152–94.

Kliemt, H. (1993), "Gerechtigkeitskriterien in der Transplantationsmedizin - eine ordoliberale Perspektive", in E. Nagl and C. Fuchs, eds., *Soziale Gerechtigkeit im Gesundheitswesen*, Berlin et al., 262–76.

Kliemt, H. (1996), "Rationierung im Gesundheitswesen als rechtsethisches Problem", in P. Oberender, ed., *Rationalisierung und Rationierung im Gesundheitswesen*, SM Verlagsgesellschaft, Gräfelfing, 23–31.

Kliemt, H. (2010), "Das Gut der Rationierung", *Zeitschrift für Wirtschaftspolitik* 59, 267–74.

Krämer, W. (1989), *Die Krankheit des Gesundheitswesens. Die Fortschrittsfalle der modernen Medizin*, 2nd ed., Frankfurt/M.

Krämer, W. (1993), *Wir kurieren uns zu Tode. Die Zukunft der modernen Medizin*, Frankfurt/M.

Loewy, E. H. (1991), "Cost Should Not Be a Factor in Medical Care", *New England Journal of Medicine* 302, 697.

Marckmann, G. and U. Siebert (2002), "Kosteneffektivität als Allokationskriterium in der Gesundheitsversorgung", *Zeitschrift für medizinische Ethik* 48,171–90.

Mechanic, D. (1992), "Professional Judgment and the Rationing of Medical Care", *University of Pennsylvania Law Review* 140, 1713–54.

Montgomery, F. U. (2011), "Ehrliche Priorisierung medizinischer Leistungen statt heimlicher Rationierung", *Interview in Forschung und Lehre* 18 (8).

Nelson, L. J. III (2011), "Rationing healthcare in Britain and the United States", *Journal of Health & Biomedical Law* 7, 175–232.

Samuelson, P. A. and W. D. Nordhaus (2001), *Economics*, 17th ed., Mc Graw-Hill, Boston.

Schultheiss, C. (2004), "Im Räderwerk impliziter Rationierung. Auswirkungen der Kostendämpfung im deutschen Gesundheitswesen", *Psychoneuro* 30, 221–26 and 568–74.

Strech, D., M. Synofzik and G. Marckmann (2008), "How Physicians Allocate Scarce Resources at the Bedside: A Systematic Review of Qualitative Studies", *Journal of Medicine and Philosophy* 33, 80–99.

Thielscher, C., T. Schüttpelz and M. Schütte (2012), „Quantification of Rationing in Germany", *Gesundheitsökonomie und Qualitätsmanagement* 17, 297–303.

Ubel, P. A. (2000), *Pricing Life. Why It's Time for Health Care Rationing*, MIT Press, Cambridge, Mass.

Walker, S., S. Palmer and M. Sculpher (2007), "The Role of NICE Technology Appraisal in NHS Rationing", *British Medical Bulletin* 81 and 82, 51–64.

Wasem, J. (2012), "Betreibt der Gemeinsame Bundesausschuss explizite Rationierung?", in Gemeinsamer Bundesausschuss, ed., *Begegnungen mit Dr. Rainer Hess*, Berlin, 202–03.

Zentrale Ethikkommission der Deutschen Ärztekammer (2000), "Prioritäten in der medizinischen Versorgung im System der Gesetzlichen Krankenversicherung (GKV): Müssen und können wir uns entscheiden?" *Deutsches Ärzteblatt* 97 (15), A-1017–A-1023.

Competition, Incentives and the English NHS

Carol Propper*

Introduction

Twenty years ago within OECD countries competition in healthcare, on either insurer or the provider side of the healthcare market, was confined to the USA. Other OECD countries operated either National Health System (NHS)-type or social insurance systems. The choice of healthcare insurer or provider was not an important component in either type of system. Choice was restricted to richer individuals in all these systems, either through a small private sector in (some of the) NHS countries, or to choice of insurance for higher income earners (for example, in the Netherlands). In the last 20 years, however, competition has been widely advocated as a reform model, either on the delivery side, or the insurance side, or on both. The UK has been a leader on the delivery side, introducing competition on the delivery side (between hospitals) in the 1990s, with the creation of the NHS internal market in 1991; and again in England in the 2000s under first the Labour administration of Tony Blair and then the current Coalition government. On the insurance side, the Netherlands has been pursuing a policy of competition since the Decker plan of the 1990s and has been actively promoting competition on the delivery side since the turn of the century. New Zealand and the Nordic countries have encouraged competition on the delivery side, while Switzerland and Germany have introduced greater competition on the insurance side.

As articulated by politicians, the appeal of competition is simple. Competition delivers greater productivity in the rest of the economy and choice is generally valued by consumers. Extending this to the healthcare sector seems a logical way of improving productivity. Competition between suppliers will encourage efficiency and raise quality, while increasing choice will meet consumer demands for a more personalised service and, in cases where there is cost sharing, it should make consumers more responsive to quality and price differences.

Yet at the same time as competition was being proposed as a reform model in Europe, the US market was consolidating, leading to a large rise in market concentration on the provider side and concerns over the operation of markets in healthcare in the USA (Gaynor and Town 2011). From other quarters, there is growing evidence of an association between volume and outcomes, particularly for high-tech services. While it is not clear whether this is due to selection or learning by doing (Gaynor and Town 2011), it has driven an interest in the consolidation of specialist services with an attendant decrease in the number of providers of these services. More generally, there is interest in the integration of primary and secondary care. All of these raise questions about the role of competition.

Gaynor and Town (2011) and Dranove (2011) provide detailed reviews of the role of competition in healthcare. Against this backdrop, the focus of this article is limited to a narrower aim: to examine what we know and don't know about competition in healthcare from reforms in the UK and England. Thus I limit my focus to competition on the provider side, as competition in insurance has not been implemented in the UK to date. The article begins with a brief summary of the main messages from the international literature on this topic, drawing heavily on Gaynor and Town (2011) before turning to the UK experience. The paper concludes by detailing the issues where very little evidence is available.

Evidence from the USA

Early research on competition followed the structure-conduct-performance (SCP) paradigm, which is theoretically underpinned by oligopoly theory. Simple models of Cournot and differentiated Bertrand competition predict a direct relationship

* University of Bristol and Imperial College London.

between market structure, firm conduct, and market performance (measured by prices and/or profits). In simpler terms, more concentrated markets facilitate behaviour that leads to higher prices and profits (Dranove 2011).

Almost all the studies are of US markets. Gaynor and Town (2011) conclude that almost all the literature finds a positive relationship between hospital concentration and price, but the strength of the relationship is affected by the structure of health insurance. Analysis of mergers (confined to US studies) supports this pro-competition conclusion. They generally show that prices increased (or increased faster relative to trend) for hospitals that consolidated relative to the control group hospitals. However, while the direction of impact of hospital mergers is clear, the estimated magnitudes are heterogeneous and vary across market settings, hospitals and insurers. There is a rapidly growing body of empirical literature on competition and quality in hospital-based healthcare. Most of the studies of Medicare patients – where prices are generally set by a regulator and individuals have close to full insurance – show a positive impact of competition on quality. This is not surprising, since economic theory for markets with regulated prices predicts such a result. However, the results from studies of markets where prices are set by firms (for example privately insured patients) are much more variable. Some studies show increased competition leading to increased quality, and some show the opposite. While this may appear surprising, it is not. Economic theory predicts that quality may either increase or decrease with increased competition when firms are determining both quality and price (Gaynor and Town 2011).

Evidence from the UK

The UK has had two periods of pro-competitive reform on the delivery side. The first was the 1990s internal market, which separated the provision of hospital care from payment for this care and allowed selective contracting between buyers and sellers of secondary healthcare. Primary care services were relatively untouched and tax-funded payments were maintained and allocated to local buyers on the basis of medical need, as before the reforms. These reforms were abandoned when the Blair administration came to power in 1997, principally due to fears of a 'two tier' system and concerns over waiting times. However, in the mid-2000s the Blair administration re-introduced competition (in England only), this time within a system of prospective payments that are very similar to the US DRG system used by Medicare. The intervening ten years had also seen the growth in information on the quality of care provided at NHS hospitals. During the 1990s no such information was publicly available. During the 2000s there has been significant growth in publicly available data on provider performance, though the data that is available to the public tends to be at a reasonably aggregate level (e.g. at a hospital, rather than an individual site level).

Evidence from the 1990s internal market

The evidence from the 1990s reforms is relatively limited, but the evidence that does exist suggests the following. Firstly, costs may have fallen more in competitive areas (Soderlund and Propper 1998). Secondly, buyers of healthcare who were primary care providers (General Practitioner (GP) fund holders) seemed to be able to extract better deals from hospitals than the larger purchasers responsible for whole populations, responsible for all the patients in their area and for purchasing emergency as well as elective care (Propper, Croxson and Shearer 2002). This was perhaps because they had stronger financial incentives, in that any gains from purchasing could be retained to put into their businesses, whilst the larger purchasers had to break even every year. The larger purchasers were also concerned about the viability of local services if they moved services at the margin, while the fund holders were less concerned with this issue as they had no remit for the provision of all secondary care services (Le Grand, Mays and Mulligan 1998). Thirdly, hospitals facing more competition focused on reducing waiting times, but at the expense of unobserved quality (Propper, Burgess and Green 2004; Propper, Burgess and Gossage 2008). The findings that waiting times fell but also unobserved quality fell, whilst uncomfortable for the proponents of competition, fall into line with the predictions from simple models of competition with imperfect information, which show that as competition increases, sellers will focus on those aspects of care for which demand is more elastic (Dranove 2011). As buyers of care during this period were interested primarily in increasing volume and reducing waiting times, and quality of care was not made public, it is not surprising that sellers engaging in competition focused on bringing down

waiting times at the expense of unmeasured quality. Fourthly, despite the political fears of two tier services, there is little evidence that patients whose secondary elective care was purchased by GP fund holders received more care than those patients covered by the larger health authorities (Cookson et al. 2010).

The evaluation of these reforms was hampered by lack of data. So for example, the most robust study of the impact of competition, which exploits pre-reform variation in hospital density, examined only waiting times and quality as measured within hospital mortality following admissions for heart attacks (Burgess et al. 2008). Whilst this measure has been used extensively in economics literature as a measure of hospital quality, death rates, whilst important, are only one aspect of quality and there are issues over their reliability when volumes of admissions are small and the measures are noisy from year to year. In addition, studies were unable to get inside the 'black box' of what exactly hospital managers and buyers were doing to bring about gains (and losses) from competition. Evaluation was also hampered by the short-lived nature of the reforms. They were only started in 1991 and ended in 1997, but even during the reform period, their effect was muted and the freedom of buyers and sellers curtailed (Le Grand et al. 1998), perhaps due to fears of the emergence of a two tier system and a more general concern on the part of central government to limit variation within the NHS.

Evidence from the English reforms of the 2000s

The reforms of the 2000s were of a similar nature to those of the 1990s, but were characterised by three important differences. Firstly, prices for elective care were set centrally using a prospective payment system similar to the US DRG system. Secondly, data on quality and other attributes of care was much more widely available. Thirdly, the incentives for sellers had been boosted through two further reforms. The first was the Foundation Trust (FT) programme. This gave hospitals deemed by the regulator to be better run greater autonomy of action, including in the retention of surpluses. Better-run status was defined primarily in terms of financial propriety and a reduction in waiting times. All hospitals could apply for FT status, so the programme essentially gave all hospitals (not just FTs) an incentive not to make losses and, possibly, to increase quality or at least not increase waiting times. The second reform was the government's promotion of entry by private sector providers supplying elective treatments for which there were long waiting lists. The evaluation of this set of reforms is ongoing, but the following stylised facts seem to be emerging.

Firstly, there is evidence that the take up of choice was slow and that GPs did not offer it to all patients (Dixon et al. 2009). Despite this, there is also evidence that patterns of care seeking changed in a manner that suggested that better quality hospitals were being chosen more often. Gaynor, Moreno-Serra and Propper (forthcoming) show that hospitals with lower pre-policy mortality rates and waiting times had a larger increase in elective patients post-policy than those with higher mortality and higher waiting times. A structural demand analysis of patients seeking elective coronary artery bypass graft treatment showed that sicker patients were more sensitive to mortality rates post-reform (Gaynor, Propper and Seiler 2012b). Secondly, two papers use the variation in the location of hospitals pre-policy to undertake a difference-in-difference analysis to derive a causal effect of competition (Cooper et al. 2011; Gaynor et al. forthcoming). They exploit the fact that hospitals located in areas where there is a higher concentration of hospitals are more exposed to the policy of competition post policy (similar to Propper et al. 2008). The papers show that death rates for patients admitted with heart attacks fell to a greater extent in hospitals located in competitive areas than in other hospitals post-policy. Gaynor et al. (forthcoming) also find that hospitals located in more competitive areas had a larger fall in mortality from all causes and lower lengths of stay for elective surgery post-policy, with no increases in overall expenditure.

The findings that quality has improved fit with the Dranove-Sattherthwaite (Dranove 2011) model of competition between hospitals. In contrast with the internal market of the 1990s, quality is better measured and price competition (at least for elective care, which was covered by the prospective payment system) was not possible. Buyers therefore care about quality and competition should increase quality. Nevertheless, the difference-in-difference approach remains open to the criticism that we don't know what is happening within the "black-box" – these papers do not present findings on how individual managers in hospitals and clinicians experienced the reforms.

One paper may shed some light on what may be driving the results. Bloom et al. (2010) examine the relationship between the quality of hospital management practices, outcomes and competition. They find that better quality management practices are associated with better NHS hospital outcomes, including lower deaths following emergency AMI (acute myocardinal infarction) admission, better financial performance, higher staff satisfaction and higher scores from the quality regulator. In addition, exploiting the fact that hospitals located in marginal political constituencies are less likely to be closed, they use political marginality to instrument the number of competitors a hospital faces. They find that competition appears to result in better management practices. As the turnover of NHS managers is high, this may be one reason why hospitals located in competitive areas have better outcomes after the reforms – as the quality of management in these hospitals is higher.

Thirdly, despite fears that poorer patients would be disadvantaged by increasing choice and competition, there seems to be little evidence that this is the case. Dixon et al. (2009) found that choice was not only exercised by the better off. Cookson, Laudicella and Li Donni (2011) also found no increase in the inequality of treatment across patients from different areas. Gaynor et al. (2012b) found that the individuals from poorer areas were more sensitive to waiting times after the reform.

The differences between the findings from the 1990s internal market and the experience of the 2000s highlight the importance of information. While the information available in the 2000s was not perfect, it was greater than in the 1990s and perhaps allowed doctors (as agents for their patients) to steer patients away from poorer performing local hospitals. The fact that prices were not part of the choice process meant that they did not have to trade off price against quality. We can say less about the importance of incentives for managers. It seems clear that achieving greater autonomy (FT status) was important for hospitals; but whether this gave them incentives to improve quality is less clear as the FT regime placed an emphasis on financial and waiting time performance rather than clinical quality.

An incomplete picture

The emerging evidence indicates that competition between hospitals can improve outcomes in an NHS setting, but unfortunately we can only see part of the picture.

- The outcomes that have been examined constitute only a small part of the whole activity of hospitals and some would argue these outcomes are not measured accurately enough to base strong conclusions upon.
- The mechanisms by which improvements have occurred are not well understood or researched.
- There are no studies of the (transactions) cost of introducing competition.
- We know little about competition in primary care settings in the UK (or elsewhere).

The drive for competition is taking place in cases where there are also calls for consolidation and vertical integration to achieve higher clinical quality. However, the evidence is limited here too.

- In a recent review of the US literature, Vogt and Town (2006) concluded that hospital market consolidations tend to increase prices, have a mixed impact on quality and achieve only modest savings, few of which are passed onto payers and consumers in terms of lower prices. A case study of a small number of hospital mergers in England concluded that these did not appear to realise large gains (Fulop et al. 2002). The scale of consolidation in England has been very large: between 1997 and 2003 approximately half of all acute hospitals were involved in a merger with other hospitals. Gaynor, Laudicella and Propper (2012a) found that these mergers reduced the volume of activity and staffing, but did not increase output per staff member and appeared to achieve no gains in terms of quality. These limited gains raise questions over a policy of unfettered mergers, as this reduces competition.
- While the model of integrated care does hold some appeal, there has been little economic analysis of this model. In a recent review, Bevan and Janus (2011) cast doubt on whether integrated care can be achieved in the UK given the historic separation of specialists within hospitals and general practitioners in the community. In its favour, integration can be achieved by contracting, as well as by the full-scale merger of primary and secondary care providers. An example is the Accountable Care Organisation

(ACO) model that is being proposed for the US system (Antos et al. 2009). There are still very few studies of integrated care organisations and of the different ways of bringing about integration, and this is likely to be a fruitful area for research.

- The (primarily medical) literature has shown a strong association between volume and better outcomes, particularly in high-tech procedures. There is some research to suggest that this is causal in some cases (Gaynor and Town 2011). If causal, then the gains from competition need to be balanced against the gains from consolidation.

References

Antos, J., J. Bertko, M. Chernew, D. Cutler, D. Goldman, M. B. McClellan, E. McGlynn, M. Pauly, L. Schaeffer and S. Shortell (2009), *Bending the Curve: Effective Steps to Address Long Term Healthcare Spending Growth*, Engleberg Centre for Healthcare Reform, Brookings, Washington D.C.

Bevan, G. and K. Janus (2011), "Why Hasn't Integrated Healthcare Developed Widely in the United States and Not at all in England?", *Journal of Health, Politics, Policy and Law* 36 (1), 141–64.

Bloom, N., C. Propper, S. Seiler and J. Van Reenen (2010), "The Impact of Competition on Management Quality: Evidence from Public Hospitals", *NBER Working Paper* no. w16032.

Cookson, R., M. Laudicella and P. Li Donni (2011), "Did Increased Competition Undermine Socio-economic Equity in Hospital Care in the English National Health Service from 2003 to 2008? Panel Analysis of Small Area Administrative Data", *Journal of Health Economics*, in press.

Cookson, R., M. Dusheiko, G. Hardman and S. Martin (2010), "Competition and Inequality: Evidence from the English National Health Service 1991-2001", *Journal of Public Administration Research and Theory* 20, 181–205.

Cooper, Z., S. Gibbons, S. Jones and A. McGuire (2011), "Does Hospital Competition Save Lives? Evidence from the English NHS Patient Choice Reforms", *Economic Journal* 121 (554), F228–F260.

Dixon, A., R. Robertson, J. Appleby, P. Burge, N. Devlin and H. Magee (2009), *Patient Choice: How Patients Choose and How Providers Respond*, Kings Fund, London.

Dranove, D. (2011), "Healthcare Markets, Regulators and Certifiers", in T. McGuire, M. Pauly and P. P. Barros, eds., *Handbook of Health Economics*, vol. 2, Elsevier/North-Holland, Amsterdam.

Fulop, N., G. Protopsaltis, A. Hutchings, A. King, P. Allen, C. Normand and R. Walters (2002), "The Process and Impact of NHS Trust Mergers: A Multi-centre Organisational Study and Management Cost Analysis", *British Medical Journal* 325, 246–49.

Gaynor, M., M. Laudicella and C. Propper (2012a), "Can Governments Do it Better? Merger Mania and Hospital Outcomes in the English NHS", *Journal of Health Economics* 31 (3), 528–43.

Gaynor, M., R. Moreno-Serra and C. Propper (forthcoming), "Death by Market Power: Reform, Competition, and Patient Outcomes in the National Health Service", *American Economic Journal: Economic Policy* (also available as *NBER Working Paper* 16164), in press.

Gaynor, M., C. Propper and S. Seiler (2012b), "Free to Choose: Reform and Demand Response in the British National Health Service, *NBER Working Paper* 18574.

Gaynor, M. and R. J. Town (2011), "Competition in Healthcare Markets", in T. McGuire, M. Pauly and P. P. Barros, eds., *Handbook of Health Economics*, vol. 2, Elsevier/North-Holland, Amsterdam.

Le Grand, J., N. Mays and J. Mulligan, eds. (1998), *Learning from the Internal Market: A Review of the Evidence*, Kings Fund, London.

Propper, C., S. Burgess and K. Green (2004), "Does Competition between Hospitals Improve the Quality of Care? Hospital Death Rates and the NHS Internal Market", *Journal of Public Economics* 88, 1247–72.

Propper, C., S. Burgess and D. Gossage (2008), "Competition and Quality: Evidence from the NHS Internal Market 1991-9", *Economic Journal* 118, 138–70.

Propper, C., B. Croxson and A. Shearer (2002), "Waiting Times for Hospital Admissions: the Impact of GP Fundholding", *Journal of Health Economics* 21, 227–52.

Soderlund, N. and C. Propper (1998), "Competition in the NHS Internal Market: An Overview of its Effects on Hospital Prices and Costs", *Health Economics* 7, 187–97.

Vogt, W., and R. Town (2006), How has Hospital Consolidation Affected the Price and Quality of Hospital Care?, *Robert Wood Johnson Foundation Synthesis Report* 9.

Competition in Health Insurance and Premium Regulation

Jan Brosse* and
Mathias Kifmann*

Introduction

In most countries, health insurance markets are highly regulated. Insurers in particular are not allowed to differentiate their premiums according to health risks and must often charge a uniform premium for all applicants. This policy is referred to as "community rating" and is used, for example, in Germany and Switzerland. It is motivated by concerns related to justice. In an unregulated market, insurers would charge those with higher health risks higher premiums, or would not even offer them coverage at all. This is regarded as unjust by many, particularly in cases where differences in health risks are beyond an individual's control.

Even if insurers are initially allowed to set risk-dependent premiums, they are often not permitted to adapt their premiums to changes in health status. Such regulation is in place in private health insurance in Germany. In the individual health insurance market in the US, most states require "guaranteed renewability" which obliges insurers to sell a contract holder a new contract with the premium at average rates for her or his initial risk class (Patel and Pauly 2002). These contracts offer insurance against "premium risk" or "reclassification risk" which arises if premiums are adapted to unforeseeable changes in the risk type.

These premium regulations have implications for the workings of health insurance markets. Some potential gains from competition are likely to be diminished. Community rating creates incentives for risk selection, which call for further regulation. Guaranteed renewability can lock-in individuals with their health insurer. Before we discuss these issues, we begin by reviewing the potential benefits and drawbacks of competition in health insurance. Our contribution will also highlight alternative policies that aim to achieve the same effects as community rating and guaranteed renewability.

Competition in health insurance: advantages and drawbacks

Competition in health insurance can yield a number of benefits for consumers. Insurers have incentives to administer contracts and to control claims efficiently in order to be able to offer contracts at prices close to the expected costs of insurance claims. Furthermore, competition encourages insurers to design insurance contracts according to individual preferences. This calls for the specification of efficient levels of co-payments taking into account the costs of insurance and moral hazard. Coverage of health services and reimbursement criteria are further dimensions of an insurance contract.[1]

Compared to other branches of insurance, health insurers can provide a range of additional services. In particular, they can act as an important agent for individuals who seek a high quality of care at reasonable prices. Under the Managed Care approach, insurers take this role by becoming organizers of healthcare. This approach contains several arrangements designed to achieve high quality and efficiency of provision, for example quality assurance and pay-for-performance programs. Measures to control healthcare expenditure often rely on restrictions of provider choice. Treatments may also need to be evaluated through "utilization reviews", and physicians can be obliged to follow special guidelines in their treatment decisions. In the special case of a Health Maintenance Organization, insurers go even further and supply services themselves by employing

* Hamburg University, Hamburg Center for Health Economics.

[1] However, insurance design can also be a means of avoiding price competition as detailed contracts may tend to confuse individuals and increase search costs (Abaluck and Gruber 2011; Schram and Sonnemans 2011).

physicians and running hospitals. So far, Managed Care is mainly used in the US. It can be found to some degree in other countries with private health insurance such as Chile and Switzerland.

The competitive pressure to bring down prices to cost, however, can also have severe drawbacks. Individuals differ substantially in their health risks and therefore in their expected healthcare expenditure. In their underwriting process, insurers usually get a good picture of the health status of an individual and adjust the premium accordingly. This leads to risk variation in premiums (an example is presented in Box 1) that is precarious in two ways. On the one hand, it is regarded as unjust, especially when individuals cannot be held responsible for their health status. On the other hand, risk rating can be disadvantageous for healthy individuals if health status changes and premiums are adjusted for new conditions. From an ex ante point of view, this generates a premium or reclassification risk to individuals, which they would like to cover by insurance. To some extent, markets can provide insurance against premium risk by offering individuals long-term coverage without individual premium adjustment. However, insurers will generally not charge uniform premiums from individuals with initial differences in health status. Empirical studies for US markets show that premiums differ considerably with respect to health risk, but also indicate some insurance of premium risk since the relationship between expected healthcare expenditure and premiums is not proportional (Pauly and Herring 1999, 2007).

A potential problem of health insurance markets is adverse selection, which arises when individuals are better informed about their health risk than insurers. However, there is little evidence relating to this phenomenon. Most health insurance markets are regulated, making it hard to distinguish between the effects of premium regulation and asymmetric information. In addition, health insurers are usually able to obtain detailed health information in the underwriting process, which limits the possible information advantage of applicants.

Box 1

Risk rating in German private health insurance

About 9 million individuals obtain their basic coverage through private health insurance (PHI) in Germany (PKV 2012). These include employees whose income exceeds a certain threshold, self-employed individuals or civil servants. Premiums in the PHI depend on initial health status. Unless major changes in overall healthcare expenditure arise, premiums must be constant throughout a lifetime. Surpluses in early years are saved to finance higher healthcare expenditure in old age. Insurers are neither allowed to terminate a contract nor may they adjust the premium to individual changes in health status.

Using their own or publicly available data, insurers calculate surcharges. For example, one German health insurer charged 50 percent extra for arthritis of one joint, 20 percent extra for allergies excluding asthma, 20 percent extra for varicose veins and 40 percent extra for the presence of gallstones. Insurers may also completely deny insurance to high-risk individuals, for example those working in dangerous occupations (for instance, lumbermen or sailors) or those who have expensive diseases such as multiple sclerosis, apoplexy and pneumoconiosis.

Community rating

The widespread regulatory response to the negative effects of risk rating is community rating, i.e. the requirement to charge uniform premiums. Sometimes this regulation is weakened by allowing premiums to vary within bands or by defining groups for which premiums can be differentiated (for example smokers vs. non-smokers).[2] At first sight, this regulation avoids unjust premium differentiation. In addition, the premium risk problem appears to be solved. However, community rating creates new challenges inducing further regulation. Firstly, low-risk individuals may find that community-rated insurance is not attractive to them. They may prefer not to buy any health insurance to avoid cross-subsidizing high risks. For this reason, community rating often goes in hand with compulsory insurance. Conversely, insurers have little incentive to insure high-risk individuals with community-rated premiums. Open enroll-

[2] In some countries, premiums also depend on income. To avoid disadvantaging insurers with low-income members, a central fund is usually introduced to correct for such differences. For example, the "*Gesundheitsfonds*" in Germany collects income-dependent contributions and pays capitations to sickness funds.

ment or guaranteed issue is therefore frequently required. A threefold regulation of community rating, open enrollment, and compulsory insurance can, for example, be found in Belgium, Germany, the Netherlands and Switzerland.

The main problem of this regulatory approach is the incentive for insurers to concentrate their efforts not on an efficient provision of services, but on risk selection because of the gap between an individual's premium and expected healthcare expenditure. Two variants of risk selection can be distinguished (Zweifel, Breyer and Kifmann 2009, 253–54). When insurers can observe characteristics of individuals related to healthcare expenditure, they can try to directly risk select by influencing contracting. For example, insurers may take their time processing the contract form handed in by a person who is predicted to be expensive. Individuals who can be considered to generate a surplus may be encouraged to sign a contract with supplementary services priced at a discount or, in extreme cases, outright payments. Indirect risk selection consists of designing benefit packages or of contracting with service providers who are attractive for low risks, but unattractive for high risks. It does not require insurers to observe risk types, but relies on self-selection since individuals with different risk types differ in their preferences.

Studies on risk selection focus on the direct variant. With a field experiment on German sickness funds, Bauhoff (2012) shows that insurers select based on geography. Individuals from West Germany have to wait longer than those from East Germany if they request a contract for membership. Direct risk selection in Germany has also been documented by the *Verbraucherzentrale*, Hamburg, a consumer advice center (see Box 2). Baumgartner and Busato (2012) investigated the extent of risk selection in Switzerland. In a field experiment, they compare insurers' reactions to young applicants willing to accept high deductibles (indicating low risks) and to old applicants preferring low co-payments (indicating high risks). They find that applicants with low risk signals have to wait about a day less for an insurer's response, are offered lower premiums and often receive offers from a subsidiary within an affiliated group, apparently specializing in low risks.

A further problem of community rating stems from the fact that low risks are more likely to switch insurers than high risks. This has been demonstrated for Germany by Nuscheler and Knaus (2005), for the Netherlands by van Vliet (2006) and for Switzerland by Beck (2004). This can threaten the existence of insurers who have a high share of high risks. If they are forced to raise the premiums, they can expect mostly low risks to leave, putting the firm in further distress.

To cope with risk selection, several measures are available. Obvious methods of direct risk selection can be legally ruled out and punished. Setting up health insurance exchanges, which make it possible to join insurers without having direct contact, may be useful. With respect to indirect risk selection, insurers can be restricted in designing their benefit packages. Minimum benefits can be defined, obliging insurers to offer benefits that are of importance to high risks, such as the treatment of chronic diseases.

Box 2

Direct risk selection in Germany

In the German sickness fund system, all funds are obliged to accept any applicant. However, they have ways of bypassing this legal requirement. This became evident in spring 2011 when one fund, CityBKK, was hit by insolvency. Its members were commonly presumed to be high risk. Members of the fund contacted other funds, some of which tried to avoid accepting their applications in the following ways:

- Funds recommended applicants to select other funds.
- One fund pointed out the disadvantages of joining it, for example that the applicant might have to take other pharmaceuticals after switching. If the applicant insisted, the employee said that his/her job would be threatened if s/he accepted former CityBKK members and hung up.
- Employees of another fund pretended that application forms had run out. The applicant was then referred to headquarters for a personal interview, but appointments were not available within the next two months.
- One fund's website was blocked in Hamburg where many CityBKK members live, making it impossible to download application forms.
- Another fund's hotline was constantly busy, and if someone could be reached, then that person pretended not to be authorized to affiliate any applicant.

Source: *Verbraucherzentrale* Hamburg 2011.

In addition, imposing an upper limit on benefits can discourage insurers from offering services that are only means to attract low risks, such as access to fitness centers (Kifmann 2002). The problem of this approach, however, is that potential benefits from health insurance competition are lost. In particular, this is evident with respect to the choice of contractual partners for the provision of services, a key element of the Managed Care approach. Clearly, this choice is a good way to attract low risks, for example by giving a large choice of specialists in athletic medicine, and to avoid high risks, for example by contracting very few experts in chronic diseases.

With risk adjustment, economists tend to favor another approach to counter risk-based selection. The objective of these schemes is to pay insurers more if they insure high risks and less if they enroll low risks. The first risk adjustment schemes relied on the easily observable characteristics of individuals such as age and gender. Meanwhile, diagnostic data is frequently used (Zweifel et al. 2009, 278–80). Without risk adjustment, the potential gains of risk selection can be large. For example, Beck, Trottmann and Zweifel (2010) examined the incentive to dump unfavorable or cream-skim favorable individuals with Swiss data. Successful dumping potentially led to a 46 percent reduction in premiums over five years in the case of no risk adjustment. This advantage fell to 16 percent if prior hospitalization and membership in a pharmacy-based cost group was added to the risk adjustment formula. Premium reductions for cream-skimming are roughly the same. Earlier studies and those of other countries also find potentially large gains, depending on the variables used for risk adjustment and the information insurers have available for risk selection (see, for example, Newhouse et al. 1989; van Barneveld et al. 2000; Shen and Ellis 2002 and Holly et al. 2003).

In European countries using community rating, the net of further regulations is tightly meshed. Insurance is compulsory and insurers must accept any applicant. Benefits are strongly regulated, even up to the point that insurers are effectively obliged to offer almost identical benefit packages as in Germany. In addition, risk adjustment schemes are in place. It is controversial whether all of these regulations are necessary. For example, with better risk adjustment in place, insurers could be given more freedom in designing their benefits and in contracting with providers. The current state, however, makes it difficult for insurers to realize the potential benefits of competition discussed above. Frequently, their role is reduced to offering a given benefit package at low cost.

If insurers were to be allowed to offer different benefit packages, for example traditional health insurance and managed care, another problem of community rating would become virulent. An efficient choice of insurance may require the relative price of these packages to depend on the risk type. With community rating, however, only one uniform price differential is possible, a priori ruling out an efficient choice (Kifmann 1999). In this environment, risk adjustment schemes that completely neutralize incentives for risk adjustment may also be impossible (Schokkaert and van de Voorde 2004).

Guaranteed renewability

Risk rating in health insurance markets also creates a challenge over the life-cycle. Unforeseeable changes in health status can cause the adaptation of premiums, thereby exposing individuals to premium risk. In Germany, regulators have responded to this risk by not permitting private health insurers to adapt their premiums to individual changes in health status. In the US, most states oblige insurers to offer guaranteed renewability: when their contract expires, policy holders must be offered a new contract with a premium at average rates for their initial risk class (Patel and Pauly 2002).

These regulations have implications for the design of insurance contracts over the life-cycle. Guaranteed renewability is mostly attractive for individuals who turn out to be high risks. Low risks, by contrast, always have the option of changing insurers. Legally, they cannot be tied to an insurer. Therefore, the problem that insurers end up with only high risks needs to be solved. With guaranteed renewable contracts, this is achieved in the form of a prepayment. Premiums at the beginning of the contract exceed current healthcare expenditure. The surplus is used to lower premiums in the future, making it attractive for both high and low risks to remain in the contract. In Germany, this goal is reached by requiring insurers to calculate premiums in a way that they remain constant over a policyholder's lifetime. Since healthcare expenditure increases with age, the premium exceeds expected costs at a young age, thus generating a prepayment.

While guaranteed renewable contracts can provide insurance against premium risk, they tie individuals strongly to their insurer since the prepayment is lost if individuals change insurer. Insurers may exploit this lock-in situation, for example by lowering the quality of their service or by trying to deny justified claims. To what extent this happens depends on the possibilities of drafting detailed contracts and on the power of reputational forces. Similar to community rating, the problem can be expected to be severer, the greater the discretion that insurers exercise in organizing healthcare. Anticipating this, individuals may be reluctant to buy insurance contracts that extend the power of insurers beyond the reimbursement of insurance claims.

An interesting question is whether guaranteed renewability needs to be mandated. In contrast to community rating, no ex ante redistribution is involved. Guaranteed renewability is only concerned with ex post changes and, therefore, risks usually covered by insurance. Markets have also provided these contracts without a requirement as in the US prior to the Health Insurance Portability and Accountability Act (Pauly and Herring 2007). A possible justification is that standardizing terms of contracts can be useful for consumers, protecting them from contracts that fail to provide substantial premium guarantees and from exploitation of the lock-in situation (Patel and Pauly 2002). In addition, guaranteed renewable contracts can protect the public from having to step in when an individual cannot afford health insurance because of a deterioration of the health status. The prepayment also provides some protection against high premiums in old age, lowering the government's need to subsidize healthcare for the elderly.

Conclusion

Premium regulation in the health insurance market is an attempt to avoid the problems generated by risk rating. Community rating tries to avoid premium differentiation, which is regarded as unjust. Guaranteed renewability is an approach to dealing with the risk of premiums being adapted to unforeseeable changes in the risk type. At first glance, these regulations are attractive for regulators because they appear to be simple and easy to implement. However, they cause a number of side-effects. With community rating the main problem is the incentive for insurers to risk select. Various additional regulations are used to minimize this problem. In particular, the enrollment process and benefit packages are regulated. Risk adjustment schemes try to compensate insurers for insuring high risks. Overall, the main problem is that these regulations can hamper the ability of insurers to offer contracts according to the preferences of individuals. Their potential to act as an organizer of medical care is severely reduced. On a smaller scale, this problem also arises with guaranteed renewable contracts. These lead to a lock-in situation with an insurer. Individuals may therefore be reluctant to give insurers too much influence over the provision of care.

Alternative solutions that try to avoid the negative consequences of risk rating and require less market intervention are therefore of interest. Pauly et al. (1992) have proposed refundable tax credits reflecting a household's risk category. Those with little or no tax liability would receive a transfer. The crucial question with this proposal is how precisely these tax credits and transfers can reflect risk types. This is also the challenge with the concept of "time-consistent health insurance" by Cochrane (1995). This alternative to guaranteed renewability envisages a separate insurance contract contingent on individuals' risk type. Individuals turning into high risks would receive an indemnity to compensate for the higher premiums of new contracts.

Zweifel and Breuer (2006) advocate premium subsidies that are paid when risk-based premiums exceed a certain percentage of household income. However, this policy creates the incentive for the insurer and the individual to include additional services in the contract to increase the transfer. Defining the benefit package in detail can be one way of avoiding this, but, as with community rating, the ability of insurance markets to offer contracts tailored to individual preferences is curtailed. Furthermore, this policy seems less suited to meet equity objectives than community rating. From a social welfare perspective, Kifmann and Roeder (2011) find that combining premium subsidies with community rating is superior for plausible correlations of health and productivity.

At the current stage, the famous trade-off between efficiency and equity seems to be unavoidable in health insurance. The potential benefits of competition in health insurance are limited by premium regulation. On the other hand, the market outcome without premium regulation is hardly acceptable for society. Advancements in risk adjustment and in tar-

geting transfers to high-risk individuals may mitigate this trade-off in the future.

References

Abaluck, J. and J. Gruber (2011), "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program", *American Economic Review* 101, 1180–210.

Bauhoff, S. (2012), "Do Health Plans Risk-Select? An Audit Study on Germany's Social Health Insurance", *Journal of Public Economics* 96, 750–59.

Baumgartner, C. and A. Busato (2012), "Risikoselektion in der Grundversicherung", *Schweizerische Ärztezeitung* 93, 510–13.

Beck, K. (2004), *Risiko Krankenversicherung: Risikomanagement in einem regulierten Krankenversicherungsmark*t, Haupt, Bern, Stuttgart, Wien.

Beck, K., M. Trottmann and P. Zweifel (2010), "Risk Adjustment in Health Insurance and its Long-term Effectiveness", *Journal of Health Economics* 29, 489–98.

Cochrane, J. H. (1995), "Time-consistent Health Insurance", *Journal of Political Economy* 103, 445–73.

Holly, A., L. Gardiol, Y. Eggli and T. Yalcin (2003), *Health-based Risk Adjustment in Switzerland: An Exploration Using Medical Information from Prior Hospitalization*, Institut d'économie et management de la santé, Lausanne, Mimeo.

Kifmann, M. (1999), "Community Rating and Choice between Traditional Health Insurance and Managed Care", *Health Economics* 8, 563–78.

Kifmann, M. (2002), "Community Rating in Health Insurance and Different Benefit Packages," *Journal of Health Economics* 21, 719–737.

Kifmann, M. and K. Roeder (2011), "Premium Subsidies and Social Health Insurance: Substitutes or Complements?", *Journal of Health Economics* 30, 1207–18.

Newhouse, J., W. Manning, E. Keeler and E. Sloss (1989), "Adjusting Capitation Rates Using Objective Health Measures and Prior Utilization", *Healthcare Financing Review* 10, 41–54.

Nuscheler, R. and T. Knaus (2005), "Risk Selection in the German Public Health Insurance System", *Health Economics* 14, 1253–71.

Patel, V. and M. V. Pauly (2002), "Guaranteed Renewability and the Problem of Risk Variation in Individual Health Insurance Markets", *Health Affairs* 21, 280–89.

Pauly, M. V., P. Danzon, P. Feldstein and J. Hoff (1992), *Responsible National Health Insurance*, AEI Press, Washington, D.C.

Pauly, M. V. and B. Herring (1999), *Pooling Health Insurance Risks*, AEI Press, Washington, D.C.

Pauly, M. V. and B. Herring (2007), "Risk Pooling and Regulation: Policy and Reality in Today's Individual Health Insurance Market", *Health Affairs* 26, 770–79.

PKV (2012), Zahlen zur Privaten Krankenversicherung, http://www.pkv.de/zahlen (accessed 30 January 2013).

Schokkaert, E. and C. van de Voorde (2004), "Risk Selection and the Specification of the Conventional Risk Adjustment Formula", *Journal of Health Economics* 23, 1237–59.

Schram, A. and J. Sonnemans (2011), "How Individuals Choose Health Insurance: An Experimental Analysis", *European Economic Review* 55, 799–819.

Shen, Y. and R. Ellis (2002), "How Profitable Is Risk Selection? A Comparison of Four Risk Adjustment Models", *Health Economics* 11, 165–174.

Van Barneveld, E., L. Lamers, R. van Vliet and W. van de Ven (2000), "Ignoring Small Predictable Profits and Losses: A New Approach for Measuring Incentives for Cream Skimming", *Healthcare Management Science* 3, 131–140.

Van Vliet, R. (2006), "Free Choice of Health Plan Combined with Risk-adjusted Capitation Payments: Are Switchers and New Enrollees Good Risks?", *Health Economics* 15, 763–74.

Verbraucherzentrale Hamburg (2011), Erfahrungen bei der Krankenkassen-Suche, http://www.vzhh.de/gesundheit/121570/erfahrungen-bei-der-krankenkassen-suche.aspx (accessed 30 January 2013).

Zweifel, P., F. Breyer and M. Kifmann (2009), *Health Economics*, Springer, Dordrecht, Heidelberg, London, New York.

Zweifel, P. and M. Breuer (2006), "The Case for Risk-based Premiums in Public Health Insurance", *Health Economics, Policy and Law* 1, 171–88.

Managed Care: Prescription for Failure? Lessons from Switzerland

Peter Zweifel*

Introduction

On 17 June 2012, a majority of 76 percent of Swiss voters said 'No' to a revision of the Federal Health Insurance Act (KVG) that would have made Managed Care (MC) the dominant variant of health insurance in Switzerland (for details, see Interpharma 2012). In economic terms, MC involves a degree of vertical integration; usually, a health insurer signs an exclusive contract with a group of physicians who commit to adhere to certain cost-saving or quality-enhancing norms. In turn, the insured are directed to this group of providers (see below for more details). The MC bill had passed Parliament in September 2011 with a comfortable majority, but initiators of a popular referendum (mainly physicians) had been able to collect the necessary 50,000 signatures (this amounts to about one percent of the country's voting population).

For many outside observers, this 'No' came as a surprise. After all, MC had been introduced in the early 1990s, even before the promulgation of the new Health Insurance Act that survived a popular referendum to become effective in 1996. After a slow start, MC picked up market share after 2005, reaching 47 percent by 2010 (Federal Office of Public Health 2011). And contrary to the United States, where major employers (who contract health insurance on behalf of their work force) had strongly promoted MC, triggering the so-called Managed Care backlash, the choice between options in basic health insurance has been a true consumer choice since 1996, with employer involvement only in supplementary coverage, if at all.

* University of Zurich.

So, how did this failure come about? This article first explains MC and then moves on to expound the potential of MC to increase the efficiency of healthcare delivery. However, MC comes at a price in that it requires consumers to give up free physician choice, while physicians have to give up payment by fee-for-service. Next, it presents experimental evidence suggesting that both Swiss general practitioners and consumers require substantial compensation to accept these restrictions. The conclusion is that the 'No' of June 2012 is due to a political failure in that politicians sought to relieve the public purse through MC without respecting citizens' preferences.

What is Managed Care?

In 1945, industrialist Henry Kaiser had construction work to do in remote areas of the north-western United States, where healthcare providers were few and far between. His solution was to hire physicians, paying them a salary for treating his workers. This solution inspired President Nixon, who in 1973 signed a law mandating major employers to include at least one so-called Health Maintenance Organization (HMO) plan in the choice of health insurance plans offered to their workers (Starr 1982). The objective was to curb the rising cost of health care impinging on the public purse, and this has remained the objective of politicians in favor of MC ever since. Accordingly, the HMO variant of MC is particularly attractive to them since it completely turns around the incentives of physicians. Earning a fixed income, (possibly augmented by a bonus if the plan makes a profit), they now have an interest in seeking out less costly treatment alternatives, rather than tending towards over-treatment. Indeed, they now would want to keep the insured healthy to begin with (hence the term 'Health Maintenance Organization'). In order to ensure that these lower-cost alternatives are, in fact, pursued by patients, members of a Managed Care Organization (MCO) are assigned to a 'gatekeeper', usually a general practitioner who is in charge of coordinating care; and referrals to a specialist or to a hospital are subject to the gatekeeper's consent.

The change of labels from HMO to MCO is not coincidental, but reflects a change of structure. HMOs met with resistance from the medical profession from the start and were denigrated as providers of low-quality care. However, they were increasingly also resisted by the insured, who resented their lock-in effect. When seriously ill, patients often preferred to be treated by a provider who did not participate in the HMO. In response to this 'HMO/MC backlash' (Marquis, Rogowski, and Escarce 2004), more flexible forms of MC were developed in the 1990s. On the demand side, some of them allow MC patients to call on outside providers on the condition that they cover the extra cost out of pocket. On the supply side, physicians formed so-called Independent Practice Associations (IPAs) designed to offer health insurers discounts on fee-for-service care. They achieve these discounts by having their members adhere to second-opinion programs in the case of referrals to specialists and hospitals, or even accept utilization review that singles out high-cost physicians for monitoring. Some of these associations negotiate directly with major employers (recall that they purchase health insurance on behalf of their workers), in which case they are called Preferred Provider Organizations (PPOs). About 90 percent of the US population is currently insured by some type of MC; however, this high share is not entirely the result of consumer choice, but also of States assigning their indigent citizens covered by the Medicaid program to MCOs in an attempt to relieve their public purse.

In the case of Switzerland, social health insurers were allowed to create HMOs starting in the early 1990s, based on a waiver of the existing law. With the Health Insurance Act of 1994, they generally obtained the right to develop new products provided that they kept their promise of lowering costs and increasing efficiency. However, the so-called any-willing-provider clause, giving all physicians the right to treat and bill patients of all Swiss social health insurers, remained in effect. With healthcare providers strongly preferring conventional fee-for-service practice (Zweifel 2011), health insurers have been facing considerable difficulties in building MCOs. Integration of the two lines of command, financial and medical, also proved to be challenging. Many of the early pioneers gave up, dissolving their MCO or selling it to a competitor. Accordingly, the market share of MCOs was slow to rise, only reaching some eight percent by 2005. More recently, nudged by continuously rising premiums, Swiss consumers have turned to MC-type contracts (mostly of the more flexible varieties rather than HMOs), pushing their present market share up to almost 50 percent (Federal Office of Public Health 2011).

How Managed Care can contribute to efficiency

According to the literature on economic policy, four properties can be used to describe the efficiency of an economy, (1) least-cost production of a given set of goods, (2) matching of production with consumer preferences, (3) dynamic efficiency, i.e. adjustment of (1) and (2) to changes in supply and demand and (4) the absence of rents that would permit producers to deviate from (1) to (3). These conditions will be applied to the healthcare sector based on the argument that this sector is part of the general economy – an argument which, of course, is very alien to healthcare professionals.

1) *Contribution of MC to least-cost production:* To the extent that fee-for-service payments contain a margin in excess of marginal cost, physicians have an incentive to sell more services than medically indicated (Zweifel and Breyer 1997, ch. 7). MC usually involves a fixed payment per year per enrolled person (a so-called capitation), which does away with this incentive. Since gatekeepers also have to cover the cost of specialist and hospital services from their capitation, they seek to reduce the number of referrals. Unlike their US counterparts, Swiss MCOs cannot negotiate special hospital rates because the Cantons (Swiss member states) are the owners of the public hospitals, which account for most of the beds. MC nevertheless holds the promise of contributing to least-cost production of healthcare services in Switzerland. This also was the main reason why its lawmakers passed the MC bill in 2011; they saw MC as a way of stabilizing health insurance contributions and of relieving both cantonal and federal budgets (note that there are means-tested subsidies for health insurance, jointly financed by the Cantons and the Confederation).

   Indeed, Lehmann and Zweifel (2004) found the short-term health care expenditure (HCE) of MC-type insurance contracts to be up to 50 percent lower than that of their fee-for-service counterparts. Panel data supplied by a major Swiss health insurer permitted to use three years of observations (1997 to 2000) to determine whether an

enrollee was above or below his or her conditional expected value of HCE. This deviation served as an indicator of latent health status in the HCE regression for the year 2000. One-third of the reduction in HCE associated with HMO-type contracts could be attributed to risk selection effects in this way, leaving two-thirds as an incentive effect. In the case of IPA-type contracts with no utilization review, only one-third of savings in HCE was attributable to changed provider incentives. While total savings were estimated to be smaller by Trottmann, Zweifel and Beck (2012), IPA-type contracts again were associated with 'true' savings amounting to one-third.

2) *Contribution of MC to closer matching of consumer preferences:* The Act of 1994 mandates cost sharing in the guise of an annual deductible, combined with a ten percent rate of coinsurance applied to total outlays exceeding the deductible. The innovation was that health insurers could offer deductibles ranging from CHF 300 to 1,200 (which have now been increased to CHF 400 to 2,500).[1] Since MC-type contracts were exempted from coinsurance, MC was a welcome addition to the menu of choices for those who are risk averse with regard to wealth, but only moderately risk averse with regard to health, causing them to accept the lock-in effect of MC. The Act of 1994 also admits bonus options for no claims in analogy to experience rating in auto insurance, for example. However, in its ordnance, government slashed an initial surcharge of ten percent on the premiums of conventional contracts for fear of bonus offers serving as an instrument of risk selection (although a risk adjustment scheme was already in place). Risk adjustment (RA) punishes an insurer (and ultimately its members) whose population consists of above-average shares of young and male individuals (for some unexpected side effects of RA, see Schoder, Sennhauser, Zweifel 2010). As a result of this surcharge, bonus options have a market share of less than one percent.

3) *Contribution to dynamic efficiency:* Whenever the set of goods and services produced is not fixed, but can be extended thanks to innovation, there is a trade-off between static and dynamic efficiency. Specifically, by granting patent protection public policy seeks to encourage dynamic efficiency; in return, the temporary monopoly enjoyed by innovators violates the 'price equals marginal cost' rule of static efficiency. In the context of medical innovation, the trade-off is slightly different, revolving around the balance between process, product, and organizational innovation (Zweifel, Breyer, and Kifmann 2009, ch. 14). Due to insurance coverage, new medical technology comes at the same (almost zero) out-of-pocket cost to patients as the older one. Therefore, patients tend to prefer (and service providers, to propose) the most advanced treatment available, without much regard for cost. Compared to product innovation, cost-reducing, process and organizational innovation have traditionally been little pursued in the healthcare sectors of industrial countries in general, and of Switzerland in particular.

The case of the canton Basel-Country is instructive. In 2009, the canton decided to upgrade and extend a hospital that was built in the 1960s, less than 5 km away from its border with the canton of Basel-City, whose several hospitals (one of them a renowned university clinic) always had sufficient capacity. More generally, Swiss public hospitals have been adopting expensive medical technology without seeking the co-operation with institutions in their vicinity, resulting in an unparalleled density of MRI and CT scanners, for example. According to Rovere and Barua (2012), there are 12.6 MRI per one million inhabitants in densely populated Switzerland, compared to 8.4 in Canada. In the case of CT scanners, the difference is even more striking, with the Swiss density at 29.6 per million, twice the value of Canada (14.6 per million). In this context, MC constitutes an organizational innovation that is expected to enhance providers' interest in process innovation (Zweifel 2005).

By redressing the balance between the three types of innovation, MC may enhance dynamic efficiency. However, when it comes to the Swiss hospital sector, its effect is limited because MCOs are not permitted to vertically integrate public hospitals. MC nevertheless serves to speed up adjustment to changes in supply and demand in ambulatory healthcare, since it creates incentives for healthcare providers to contribute to the success of the health insurance plan. MC therefore injects 'competition between systems' into the healthcare sector, causing providers to be more responsive to the changing preferences of their clientele lest the

[1] CHF (Swiss franc) equals roughly 0.9 USD at current exchange rates; for more detail on Swiss health insurance, see Kreier and Zweifel 2011.

MCO lose market share, while making insurers more prudent purchasers of healthcare services.

4) *Ensuring the 'no rents' condition:* Producers who enjoy rents have little incentive to comply with criteria (1) to (3). They have leeway to deviate from least-cost production, do not have to closely match the goods and services offered to consumer preferences, and need not strive to adjust to ever-changing demand and supply conditions in order to survive in the market. Barriers to entry are known to create this leeway. Clearly, barriers to entry are as prominent in Swiss healthcare markets as those in any other industrial country (see, for example, Götte and Hammes 1998). While MC cannot do away with barriers to entry, it does establish a benchmark against which established service providers can be measured.

Warnings from experimental evidence

As noted above, the market share of MC remained low in Switzerland well past 2000, giving rise to the suspicion that MC did not conform to average Swiss preferences. Therefore, a so-called discrete choice experiment (DCE) involving some 1,000 residents was conducted in 2003. DCEs are a tool for measuring preferences for goods that are not (yet) on the market; in this present case these are MC contracts that would better match consumers' preferences. Usually, the status quo is fixed in terms of a set of attributes, while several alternatives with changed levels of these attributes are proposed to participants in the experiment, who have to indicate whether they want to stay with the status quo or whether they prefer the alternative. By making the price to be paid one attribute, one can infer the (marginal) willingness to pay (WTP) for an attribute using econometric methods (see, for example, Louviere, Hensher and Swait 2000). In the present context, the attributes were (1) physician choice (free in the status quo vs. constrained under MC), (2) access to newest medical technology (immediate vs. delayed by two years under MC), (3) coverage of pharmaceuticals (unconstrained vs. generics or cheapest alternative available), (4) Drugs for minor complaints (unconstrained vs. exclusion from the drug benefit), (4) hospital choice (community hospitals vs. regional health centers under MC), and (5) annual contribution to social health insurance (unchanged vs. up to +/- 50 CHF per month). It is worth noting that the variation in attribute (5) may not be realistic given an average contribution of CHF 3,600 (300 per month) at the time; however, it serves to move respondents back and forth between the status quo and the alternatives. If they stay with the status quo, little can be learned about their preferences.

The estimated WTP values are displayed in Table 1 (see also Zweifel, Telser, Vaterlaus 2006). They are all negative, implying that on average, Swiss consumers need to be compensated to accept the restrictions in choice imposed on them by MC. Giving up free physician choice (a defining characteristic of MC) would have to be compensated by up to 38 percent of average premium (amounting to some EUR 2,030 per year). MC could also be used to direct patients to

Table 1

Willingness-to-pay (WTP) values for MC-type attributes

WTP in € / year 1 € = 1.55 CHF (2003)	WTP Switzerland	WTP Germany	WTP Netherlands
Physician list, cost criteria only	-792*	n.a.	n.a.
Physician list, quality criteria	-408*	n.a.	n.a.
Physician list, both cost and quality criteria	-324*	-115*	-346*
Delay of 2 years in the introduction of new therapies	-503*	n.a.	n.a.
Generica only in drug benefit	-23	n.a.	n.a.
Exclusion of drugs for minor complaint from drug benefit	-46	n.a.	n.a.
Regional hospital units only	-286*	n.a.	n.a.
Constant	-451*	-500*	-256*

Notes: Figures for Switzerland refer to 2003, for Germany to 2005, and for the Netherlands to 2006.
* indicates significance at the 5 percent level or better. The WTP of -346 is the estimated WTP of the Dutch to revert from gatekeeping to free physician choice.
Source: The author.

hospitals providing better care at lower cost, in keeping with established medical opinion claiming that larger hospitals achieve better quality of treatment because of their higher volumes of surgery (Birkmeyer, Siewers and Finlayson 2002). However, this view is refuted by consumers; indeed, such a concentration would have to be compensated for by about 18 percent of average premium. In hindsight, a preference for community hospitals is not so astonishing. It suffices to imagine a future mother considering having her baby 50 km away from home, in one of these 'efficient' specialized units. Would her husband, her relatives, and her friends be likely to show up with that bunch of flowers?

One could argue that, in spite of the detailed scenario description in the DCE, most participants did not understand what MC meant, since its market share was still low in 2003. However, evidence from the Netherlands suggests otherwise. In another DCE fielded in Germany and in the Netherlands (where gatekeeping is part of the status quo), substantial WTP for returning to the status quo prior to free physician choice was found (MacNeil Vroomen and Zweifel 2011; Zweifel, Rischatsch and Leukert 2010). Interestingly, this WTP value even exceeds the compensation requested by German participants for moving away from their status quo of free physician choice and towards gatekeeping (see Table 1).

A case of political failure

The MC bill as passed by the Swiss parliament was inconsistent from the outset, because it contained a provision to force social health insurers to create MC plans everywhere – even in a canton like Uri. This canton, situated in the valley leading up to the Gotthard pass, has a population of 30,000, living at 1,400m altitude and higher, hours away from the hospital of the capital town Altdorf when the freeway is clogged by vacationers on their way to Italy. At the same time, physicians were to retain the right to conventional fee-for-service practice, which they strongly prefer (Rischatsch and Zweifel 2012).

With the market share of MC increasing anyway, Swiss politicians could abstain from nudging consumers towards MC. However, the promise of savings (to the public purse, of course) is too much of a lure to them. In its fall 2012 session, Switzerland's federal parliament already came up with a new, less restrictive MC bill. It remains to be seen whether this bill will be challenged again by a popular referendum; and if so, whether it will survive the challenge.

References

Birkmeyer J. D., A. E. Siewers, E. V. A. Finlayson, T. A. Stukel, F. L. Lucas, I. Batista, H. G. Welch and D. E. Wennberg (2002), "Hospital Volume and Surgical Mortality in the United States", *The New England Journal of Medicine* 346 (15), 1128–37.

Federal Office of Public Health (2011), Statistik der Obligatorischen Krankenversicherung 2009 (Statistics on Mandatory Health Insurance 2009), www.bag.amdin.ch/shop/00102/index.html?lang=de (accessed 14 November 2012).

Götte L. and K. Hammes (1998), "Physician Licensing", in P. Zweifel, L. Söderström and C. H. Lytkens, eds., *Regulation of Health: Case Studies of Sweden and Switzerland,* Kluwer, Boston, ch. 4.

Interpharma (2012), Managed Care (in German and French), www.interpharma.ch/de/politik/Managed-Care.asp (accessed 22 November 2012).

Kreier R. and P. Zweifel (2011), "Health Insurance in Switzerland: A Closer Look at a System Often Offered as a Model for the United States", *Hofstra Law Review* 39 (89), 89–110.

Louviere J. J., D. A. Hensher and J. D. Swait (2000), *Stated Choice Methods – Analysis and Applications,* Cambridge University Press, Cambridge, MA.

Vroomen, J. M. and P. Zweifel (2011), "Preferences for Health Insurance and Health Status: Does it Matter Whether You Are Dutch or German?" *European Journal of Health Economics* 12 (1), 87–95.

Marquis S. M., J. A. Rogowski and J. J. Escarce (2004), "The Managed Care Backlash: Did Consumers Vote with Their Feet?" *Inquiry* 4, 376–90.

Rischatsch M. and P. Zweifel (2012), "What Do Physicians Dislike about Managed Care? Evidence from a Choice Experiment", *European Journal of Health Economics*, published online 21 June 2012.

Rovere M. and B. Barua (2012), "Opportunity for Health Reform: Lessons from Switzerland", Fraser Forum July/August 2012, Fraser Institute, Canada.

Schoder J., M. Sennhauser and P. Zweifel (2010), "Fine Tuning of Health Regulation: Unhealthy Consequences for an Individual Insurer", *International Journal of the Economics of Business* 17 (3), 21–31.

Starr P. (1982), *The Social Transformation of American Medicine: The Rise of a Sovereign Profession and the Making of a Vast Industry*, Basic Books, New York.

Trottmann M., P. Zweifel and K. Beck (2012), "Demand-side and Supply-side Cost Sharing in Deregulated Health Insurance: Which is More Effective?", *Journal of Health Economics* 31 (1), 231–42.

Zweifel P. (2005), "Diffusion of Hospital Innovations in Different Institutional Settings", *International Journal of the Economics of Business* 2 (3), 465–83.

Zweifel P. (2011), "Swiss Experiment Shows Physicians, Consumers Want Significant Compensation to Embrace Coordinated Care", *Health Affairs* 30 (3), 510–15.

Zweifel P. and F. Breyer (1997), *Health Economics*, Oxford University Press, New York.

Zweifel P., F. Breyer and M. Kifmann (2009), *Health Economics*, Springer, Boston.

Zweifel P., M. Rischatsch and K. Leukert (2010), "Preferences for Health Insurance in Germany and the Netherlans – A Tale of Two Countries", *Working Paper* 1001, Socioeconomic Institute, University of Zurich.

Zweifel P., H. Telser and S. Vaterlaus (2006), "Consumer resistance against Regulation: The Case of Health Care", *Journal of Regulatory Economics* 29 (3), 319–32.

Concerns over the Financial Sustainability of the Dutch Healthcare System

Hans Maarse*,
Patrick Jeurissen** and
Dirk Ruwaard*

Introduction

The most pressing problem in current Dutch healthcare is how to guarantee its financial sustainability in the future (Ruwaard 2012). In the period 2001–2010 the real growth in healthcare expenditure averaged at 4.4 percent a year, compared to 2.2 percent in the period 1981–2000, while healthcare as a percentage of GDP peaked at 13.2 percent in 2010 (CPB 2011). With USD 5,056 per capita the Netherlands was the third-largest spender on healthcare in Europe in 2010; topped only by Norway (USD 5,388) and Switzerland (USD 5,270) (OECD 2012). Depending on the assumptions made, healthcare is projected to consume between 22–31 percent of GDP in 2040 (CPB 2011). The big political and social challenge is how to rein in the growth of healthcare expenditure without compromising the principles of universal access, solidarity and quality of care (Maarse 2011). This article gives a brief overview of some recent developments in Dutch healthcare and reforms to address the sustainability problem.

Health insurance

After almost two decades of political debate the Health Insurance Act (*Zorgverzekeringswet*) came into force in 2006. The new legislation introduced a single mandatory *basic* health insurance scheme covering the entire population. The regulatory framework encourages competition among insurers and providers, but simultaneously respects the legacy of the past by upholding the principles of solidarity and universal access. The legislation obliges each citizen to purchase a basic health plan covering, among others, family medicine, maternity care, pharmaceuticals and hospital care. There is open enrolment and citizens may switch to another insurer or health plan at the end of each year. Insurers compete on their nominal premium rate which averaged at EUR 1,361 in 2012 (NZa 2012a). Insurers are required to apply community rating: any form of experience-rating is forbidden. People on low income are compensated by a tax credit system to limit the premium that they pay to five percent of their income. Those insured also pay an income-related contribution through their employer (7.75 percent over a maximum of EUR 51,000). Furthermore, the state pays the premium for children under 18. To prevent risk selection and to achieve a level-playing field, a sophisticated risk equalisation mechanism is in place to level off differences between the insurers' risk profile. The mandatory deductible, introduced in 2008 after the failure of the no claim regime, doubled from EUR 170 a person in 2008 to EUR 350 in 2013. The costs of General Practitioner (GP) consultations, maternity care and healthcare to children under 18 are exempted from the mandatory deductible.

Insurers not only compete in basic health insurance, but also in *complementary* health insurance, where they are free to apply experience-based and medical underwriting. However, they have largely abstained from using these instruments to date. They can also make up their benefit package. Complementary plans cover extra services (for example dental care, physiotherapy). Contrary to basic insurance, complementary insurance is voluntary. The percentage of people without a complementary plan is still high but has fallen from 94 percent in 2012 to 88 percent in 2012 (NZa 2012a).

Consumer mobility peaked at 18 percent in 2006, but fell back to 3.6 percent in 2009 (NZa 2012a). Since 2010 it has gradually increased to an estimated 7.5 percent in 2013, highlighting the fact that competition has intensified.

* University of Maastricht.
** Ministry of Public Health, Welfare and Sport.

Whereas the regulator set the minimum solvency rate of insurers at 11 percent in 2012, it averaged at about 18 percent in 2011. Administrative costs are only 4.3 percent of total premium revenues (NZa 2012a). The insurers' sound financial record was partly the result of various safety nets in health insurance to give insurers some financial protection. These nets were largely abolished in 2012 to encourage insurers to engage in efficient contracting with healthcare providers. At present insurers are at risk for 91 percent of their expenses compared to 23 percent in 2006.

Efficient contracting by health insurers plays an important role in the government's policy to attain financial sustainability. Other policy measures under discussion include a further raise of the mandatory deductible, the introduction of a co-payment regime and a critical assessment of the basic benefit package. However, political and popular support for these measures is low. An interesting question is how complementary health insurance will develop. Package reduction in basic health insurance tends to be followed by an extension of complementary health insurance. This process could result in the substantial growth of a new private insurance market next to the market for statutory (basic) health insurance. In 2011 premium revenues from complementary plans already amounted to 13.6 percent of total premium revenues from basic health insurance in 2011 (Vektis 2012).

Healthcare provision

Healthcare provision has undergone several changes over the last decade. Traditionally, general practitioners fulfil a gatekeeper role. In 2011 the number of referrals to a medical specialist per 1,000 registered patients was 199 (NZa 2012b). Various initiatives are being taken to reinforce the pivotal role of general practitioners and, where possible, to reduce the number of referrals. If successful, these initiatives could save money and contribute to financial sustainability.

A noteworthy development is the introduction of integrated care pathways for patients with chronic disease (for example diabetes, COPD and vascular risk management), which is supported by a bundled payment model to pay for the providers involved in the care pathway. Insurers negotiate with the organisations coordinating the care pathway on an overall (bundled) tariff per patient. From a cost saving perspective, the model of integrated care pathways has failed so far (Struijs, van Til and Baan 2011).

The introduction of care pathways fits into a broader trend towards getting more value for money. Provider associations, patient organisations, the Healthcare Inspectorate and other stakeholders are devoting a lot of energy to the development of quality guidelines and quality measurement by means of health outcome and other indicators. Public reporting on quality is assumed to stimulate providers to perform better and care users to be more critical. Some experts see plenty of opportunity to perform better while *lowering* costs (Porter and Olmstedt Teisberg 2006), among others, by cutting the link between volume and revenues and by encouraging providers to spend more time on discussing treatment options with their patients. Quality-based funding and shared decision-making can save costs (Klink 2012).

Concentration and specialisation point to another new development in quality management. To optimally benefit from the quality-volume spiral, several complex surgical procedures are now being concentrated in a limited number of hospitals. Many hospitals can no longer meet the quality standards set by the respective medical communities and insurers are increasingly unwilling to contract each hospital for the entire spectrum of medical care. How this development will further unfold in the future, is difficult to predict. Many experts believe that the number of general hospitals will fall significantly and that a new relationship will evolve between top-clinical centres and outreach hospital facilities.

The market reform includes a significant deregulation of state hospital planning. Hospitals have become free to decide on their specialty portfolio, bed capacity, capital investments, and other issues. The extension of discretionary power was paralleled by a prospective payment model that increased their financial risk. It is assumed that this reform will improve allocation and save costs, because providers must now take the financial risks of their expansion into full account. A realistic business plan has become an indispensable instrument in provider management.

All general hospitals are private organisations, but health legislation still includes a ban on for-profit hospital care (Jeurissen 2010). If a hospital manages

to realise a budget surplus, it can either reinvest the surplus or add it to its financial reserves. This arrangement also applies to the two hospitals which are presently owned by a commercial corporation. The government recently announced a plan to lift the ban on for-profit hospital care. However, the new regulation will feature strict conditions to keep 'unwelcome' investors outside and prevent hospitals from becoming profit-maximising agencies. For-profit hospital care is still a politically sensitive topic in Dutch healthcare.

The market reform has induced an explosion in the number of independent treatment centres. Many small-scale centres have entered the market, most of which provide routine elective care in various specialty areas such as ophthalmology, orthopaedic surgery, dermatology, radiology, and many others. The number of centres, many of which are (co)-owned by hospitals, rose from 30 in 2000 to about 180 in 2011, which is almost twice as high as the number of hospitals. Despite this rapid growth, the revenues of the centres have remained limited to 3.5 percent of total expenditure for hospital care (Boer & Croon 2011).

Contracting (global budgeting)

A cornerstone of the market reform is that insurers contract efficiently. A recent report (Significant 2012) highlighted various initiatives, but the overall picture is that efficient contracting is still at an early stage. This is also true for efficient contracting by means of selective contracting. Reasons why selective contracting has remained restricted to date include a lack of information on costs and quality, market structure, the absence of powerful incentives due to safety nets, and the insurers' fear of damage among customers. Selective contracting has only been applied for some independent treatment centres and some specific medical treatments (e.g. breast cancer surgery).

The scope of free-pricing should not be overstated. In hospital care it has gradually been extended from ten percent in 2005 to 20 percent in 2008, 34 percent in 2009 and 70 percent in 2012. The prices of the rest of hospital care, as well as the fees charged by the self-employed specialists are centrally regulated by the Healthcare Authority. With some exceptions, insurers have largely abided by collective price-setting for general practitioners, physiotherapists and other providers to date. A notable event took place in 2012 when the tariffs of dental care were liberalised. Due to significant price increases the Minister of Health was forced, under heavy political pressure, to repeal the liberalisation only a few months after it had been introduced.

According to the Dutch Healthcare Authority, net prices in the liberalised hospital sector have declined relative to the regulated sector (NZa 2012c), but this conclusion has been disputed (Van der Meulen and van der Kwartel 2012). Nevertheless, the overall picture is mixed because total hospital revenues increased by an average of 6.2 percent a year in the period 2006–2010. The most important explanation of this increase seems to be a changing treatment pattern: faster active intervention, more interventions per patient and the introduction of new, more costly interventions (NZa 2012d). As regards pharmaceutical care, competition has been successful. By requiring doctors to prescribe, where possible, generics and reimbursing only the costs of low-priced generic drugs, insurers managed to implement substantial price cuts, which for some drugs even totalled 90 percent.

To control healthcare costs, the Minister of Health does not fully rely on the effects of competition. As a last resort, the instrument of budget control has remained available. Each year the minister sets a macro-budget for hospital care (and other sectors) that may not be overrun. When there is an overrun, hospitals are required to pay back the amount of overspending. Partly to avoid this unpopular measure, the minister signed a covenant with the hospital sector and health insurers in 2011 whereby the participants agreed to limit the volume growth to a maximum of 2.5 percent a year. In 2011 the Minister also signed a covenant with the association of medical specialists on the re-introduction of a macro-budget, after the lifting of a similar regime had been followed by a cost explosion. The covenant also contains substantial tariff cuts to undo the cost explosion. The use of covenants demonstrates the hybrid character of competition: market regulation is complemented with a classic form of corporatist governance.

Long-term care

The rapid growth of expenditure on long-term care (LTC) is seen as a serious threat to the future sus-

tainability of healthcare. In the period 1998–2010 public expenditure on LTC as percentage of GDP grew from 3.1 percent in 1998 to 4.3 percent (CPB 2012) and this percentage is expected to rise to 7–9 percent in 2040, depending on the assumptions made (CPB 2011). A recent OECD-report found, that in Europe only Sweden spends a higher percentage of its GDP on LTC (OECD 2011).

LTC is known as a well-developed part of Dutch healthcare. It is shaped as a mainly publicly-funded service delivered by private not-for-profit providers. The Exceptional Medical Expenses Act (AWBZ), in place since 1968, covers the bulk of expenditure, and is a truly national and largely contribution-based scheme that pays for the costs of residential care and all kinds of outpatient and home-based services for the elderly, the disabled and other categories of vulnerable people. The share of co-payments for inpatient LTC dropped from 8.8 percent in 2002 to 7.2 percent in 2011. Most clients apply for care-in-kind, but since the mid-1990s they have also been able to apply for a personal budget to purchase health services privately. The cost explosion of the personal budget scheme from EUR 413 million in 2002 to 2.3 billion in 2010 (Sadiray et al. 2011) highlights its popularity. However, experts worry that it did not lower the demand for in-kind care and also tend to crowd out informal care. Another arrangement is the Social Support Act (Wmo), in place since 2007, which pays, amongst other things, for domiciliary care. Municipalities receive a state grant to provide services which were previously covered by the AWBZ.

The ageing of the population is only one factor explaining the expenditure growth. Other factors include the government's priority around the year 2000 to reduce waiting times to socially acceptable lengths, the ambiguous description of entitlements and, consequently, the rather generous structure of the benefit package. An alarming result of recent analyses is that a substantial portion of the cost increase can be explained by the growth of less severe cases receiving LTC-services.

In recent years the government took several measures to slow down the growth of expenditure on LTC, in particular by removing some personal assistance services from the AWBZ-package and reintroducing a pseudo-budget system. For the next four-year period other substantial retrenchment programs have been announced, especially day care provisions and domiciliary services. Another measure is to upgrade the role of municipalities in LTC with the transition of domiciliary services from the AWBZ to the Wmo as prime example. Policymakers assume that local government is best informed about the local situation and also in the best position to deliver efficient, client-centred and integrated support to LTC-clients because of its responsibility for various adjacent policy areas including housing, welfare programmes, transport and local planning. Whereas competition has remained minimal under the AWBZ, municipalities have made use of competitive bidding and other strategies to cut prices. Domiciliary services have become one of the most competitive areas in healthcare. Presumably the most controversial proposal was to implement a substantial retrenchment of the personal budget arrangement whereby only a small percentage of clients would retain the option of a personal budget. Not surprisingly, the proposal was heavily disputed and when the government fell in 2012, the political crisis was immediately seized as an opportunity for mitigation.

On a more fundamental level, the government also sought to initiate a debate over individual responsibility for LTC. In its view individual responsibility has to be reinforced to keep LTC accessible to those who really need it. Each person should live as long as possible autonomously in his or her own environment and the use of intramural services needs to be scaled down. However, reinforcing individual responsibility is not only an ambiguous concept, but also a controversial strategy that keeps parties divided.

One element stands out in the political debate, however, and that is the future of the AWBZ. In its present form, it covers a wide range in inpatient and outpatient services. An important policy issue is to reform the AWBZ in accordance with its original objective: a scheme to cover the costs of people in need of long-term care (mainly people with a serious physical or mental disability). For these categories some form of social insurance scheme should remain in place. All other services must be 'delisted' and accommodated in a *provision-based* scheme. Not surprisingly, this is a politically sensitive issue.

Future perspectives

The sustainability of healthcare is a good example of what policy analysts call a wicked or unstructured

policy problem. There is little consensus on the objectives of healthcare policymaking. Opinions on how much a nation should spend on healthcare and how to translate principles as universal access and solidarity into concrete arrangements differ widely. Neither is there consensus on the instruments to achieve these objectives. One may speak of an ongoing ideological controversy, which is exacerbated by the fact that the acceptance of 'evidence' is strongly influenced by one's ideas about what a fair healthcare system should look like. At the same time the demand for healthcare continues to rise and new, often costly, interventions will become available.

It is evident that new approaches are needed to achieve the 'triple aim': better population health and higher quality for lower costs (Berwick, Nolan and Whittington 2008). However, these approaches are not easy to put into practice. For instance, there is a great need for effective prevention, but prevention may raise complex questions about individual freedom and costs. Another urgent issue is to shift the focus from health volume towards health outcome. Unfortunately, institutionalised patterns often work as a formidable barrier to change and 'best practices' do not spread quickly. Many possibilities to get more value for less money have remained unexploited yet. There is also a great need for more individual responsibility: universal access and solidarity cannot be upheld without more emphasis on individual responsibility. However, the practical implications and public acceptance of more individual responsibility appear troublesome.

Present healthcare faces a prisoner's dilemma. All players have a common interest in hard measures to guarantee its future sustainability, but none of them has an individual interest to give in and, hence, look at the other. Without political imagination and courage, the inevitable result will be paralysis in which, ultimately, all players are not only collectively, but also individually worse off.

References

Berwick, D., T. Nolan and J. Whittington (2008), "The Triple Aim: Care, Health, and Cost", *Health Affairs* 27, 759–69.

Boer & Croon (2011), *Zelfstandige behandelcentra: Kwaliteit van zorg, efficiëntie en innovatiekracht*, Naarden.

CPB (Centraal Planbureau) (2011), *Financiering onder druk*, Den Haag.

CPB (2012), *Macro Economische Verkenning* 2012, Den Haag.

Jeurissen, P. (2010), *For-profit Hospitals: A Comparative and Longitudinal Study of the For-profit Hospital Sector in Four Western Countries*, dissertation, Rotterdam.

Klink, A. (2012), *Toerusting in de arena van de gezondheidszorg*, inaugural speech, Amsterdam.

Maarse, J. (2011), *Markthervorming in de zorg*, Datawyse, Maastricht.

NZa (Nederlandse Zorgautoriteit) (2012a), *Marktscan zorgverzekeringsmarkt* 2012, Utrecht.

NZa (2012b), *Marktscan huisartsenzorg*, Utrecht.

NZa (2012c), *Marktscan medisch-specialistische zorg* 2012, Utrecht.

NZa (2012d), *Marktscan medisch specialistische zorg*: Weergave van de markt 2008-2012, Utrecht.

OECD (2011), *Help Wanted: Providing and Paying for Long-term Care*, OECD Publishing, Paris.

OECD (2012), *OECD Health Data 2012*, OECD Publishing, Paris.

Porter, M. and E. Olmstedt Teisberg (2006), *Redefining Health Care: Creating Value-based Competition on Results,* Harvard Business School Press, Boston.

Ruwaard, D. (2012), *De weg van nazorg naar voorzorg: buiten de gebaande paden*, inaugural speech, Océ Business Services, Maastricht.

Sadiray, K., D. Oudijk, H. van Kempen and J. Stevens (2011), *De opmars van het PGB*, Sociaal en Cultureel Planbureau, Den Haag.

Significant (2012), *Doelmatigheid in de zorginkoop*, Barneveld.

Struijs, J., J. van Til and C. Baan (2010), *Experimenting with Bundled Payments for Diabetes Care in the Netherlands: The First Tangible Effects*, RIVM, Bilthoven.

Van der Meulen, L. and A. van der Kwartel (2012), *Sturen op doelmatigheid,* KIWA, Utrecht.

Vektis (2012), *Zorgverzekeraars- en financiering*, Zeist.

Should Doctors Run Hospitals?

Amanda Goodall*

Introduction

The question of whether hospitals are better run by doctors or non-medically trained managers has been hotly debated for a number of years. In the past, hospitals were routinely led by doctors. All that has changed. In the UK and the US, most hospital chief executive officers (CEOs) are now non-physician managers rather than physicians (Falcone and Satiani 2008). Of the 6,500 hospitals in the US, only 235 are led by physicians (Gunderman and Kanter 2009).

It has been suggested that placing physicians in leadership positions can result in improved hospital performance and patient care (Horton 2008, Falcone and Satiani 2008, Darzi 2009, Candace and Giordana 2009, Dwyer 2010). A few years ago the UK established five academic health science centres. Their mission is to bring the practice of medicine closer to research – in the hope that innovative science can be more quickly translated into clinical procedures (Smith 2009). Physician leadership was also prioritised in the 2008 National Health Service (NHS) review (Darzi 2008, 2009). Some outstanding US medical facilities – for example the Cleveland and Mayo clinics – have explicitly introduced leadership training (for example, Stoller, Berkowitz and Bailin 2007, Stoller 2013), and management and leadership education is being incorporated into medical degrees.

Despite the growing body of research into hospital performance, there are currently no empirical studies that assess the physician-leadership hypothesis that hospitals perform better when they are led by doctors. To establish a clear relationship between leadership and organisational outcomes is challenging. Unlike in medical trials, random assignment – in this case of chief executive officers to hospitals – cannot be used. My research provides an empirical inquiry (Goodall 2011). It looks at the leaders currently being hired by hospitals and examines whether CEOs in hospitals ranked higher are typically physicians or non-medical managers.

Specialist leaders versus generalists

The issue about whether hospital leaders are, or should be, doctors or managers relates to the larger question about specialist leaders versus generalists. This topic is germane because there is recent evidence that major US firms have moved away from hiring CEOs who are specialists and towards the selection of generalist leaders (Frydman 2007; Bertrand 2009). Frydman (2007) examines the career paths of the three highest-paid executives from 1936 to 2003 (total of 708 managers) in the top fifty US public corporations (in the year 1960). She patterns a rise in the number of business degrees held by executives, and a concomitant decline in technical degrees (science, engineering and law). As the overwhelming majority of hospital leaders in the US are general managers (85 percent), it seems likely that hospital management has followed this same trend.

I first considered the question of specialist leaders versus generalists in the context of research universities, after having worked closely with two organisational leaders. I noticed that both presidents that I worked with had different ideas about institutional priorities: one, who had been an obsessive and very highly cited researcher, focused on hiring great scholars; whereas the other, who had stopped doing research early in his career to become an administrator, seemed less interested in research output and scholarship. This led me to ask the question, who should lead research universities? Should they essentially be good scholars or good managers?

* Cass Business School London / IZA Bonn.

My study of university presidents was published in several journals and a book (Goodall 2009b). The findings suggests that there is a relationship between university performance and leadership by an accomplished scholar (Goodall 2006; 2009a,b). I found that not only were the best universities in the world more likely to be led by outstanding scholars (e.g. the Stanfords and MITs), but I could also show, in longitudinal data, that universities improved their performance over time when better scholars took the reins. Thus, I found that a leader's characteristics (success in scholarship) were closely aligned with the core business activity of a university (research and teaching).

Over the last few years I have examined the question of how much core business knowledge leaders should have in a number of different settings. One was the highly-skilled environment of basketball, where it is possible to clearly identify the coaches' characteristics and teams' performance. In a study with Larry Kahn and Andrew Oswald we found a strong relationship between brilliance as a basketball player and the (much later) winning percentage and playoff success of that person as a basketball coach. Indeed, we found that the better the player (they played for the All-Stars), the better their performance as a coach (Goodall, Kahn and Oswald, 2011). In my most recent study I have shifted setting again, this time looking at the competitive industry of Formula 1 World Constructors' Championship. My co-author Ganna Pogrebna and I use six decades of field data from Formula 1. In our study we measure the change in leader (F1 principal), with the change in performance (the number of Grand Prix wins and podiums) over the 60 years. In our calculations we control for the race circuit, the race year, the constructors (McLaren, Red Bull, Ferrari, etc), and the number of cars that qualified. Our primary results show that the most successful team leaders in Formula 1 motor racing are more likely to have started their careers as drivers or mechanics – as compared with leaders who were principally managers or engineers (with degrees). When we looked further into the data we found that the result is driven by team principals who were themselves former racing drivers. In other words, time spent as a driver has a big effect on future performance as a leader. The extra probability of gaining a podium position when a driver has had a decade's experience of competitive racing is about one-in-seven.

Studying CEOs of top-ranked US hospitals

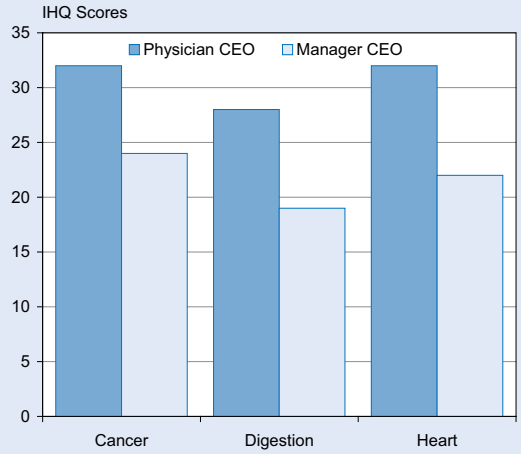
The study of hospital leaders outlined here uses a simple cross-section methodology at a given time. It therefore cannot make claims about the effectiveness of leaders; instead, it can shed light on who top hospitals hire as CEO. The wealthiest and most prestigious hospitals arguably have the widest choice of leadership candidates. If it can be shown that hospitals positioned higher in a widely-used media ranking are more likely to be led by medical experts rather than managers, this is one form of evidence that physician-leaders may make effective CEOs.

The paper identifies the CEOs in the top ranked hospitals in America – determining whether those hospitals situated higher in the league-table are more likely to be headed by physician-leaders or by professional managers. To this end, one particular quality ranking is used, namely the league tables produced by US News and World Report's "Best Hospitals" 2009.

The US News and World Report ranking is designed to inform consumers about where to seek treatments for serious or complex medical problems. Media-generated league tables cannot be viewed as entirely reliable measures of quality. However, using rating systems as heuristic devices to assess healthcare providers has nonetheless become common in the US (Schneider and Epstein 1998) and it has been shown to influence consumers' behaviour (Pope 2009). I use this ranking because it is one of the most

Figure 1

Mean Index of Hospital Quality (IHQ)
Score of hospitals led by physician CEOs and manager CEOs in three specialty fields



Source: The author.

well-established in its field. The dataset in my study covers the top-100 hospitals in the three specialist fields of cancer, digestive disorders and heart and heart surgery. Each hospital CEO is then identified and classified into one of two categories – physician-leaders, who have been trained in medicine (MD), and leaders who are non-physician managers.

Physician-led hospitals are higher-quality hospitals

To establish whether hospitals higher in the rankings are more likely to be led by physicians, I use t-tests and regression equations. I do this for the top-100 hospitals in each of the three medical fields of cancer, heart and heart surgery and digestive disorders.

In the field of cancer there are 51 physician-leaders among this set of 100 CEOs. Thirty-three are in the top-50 hospitals and 18 lead hospitals in the lower 50 group. For the other two specialities, there are 34 physician-leaders in the top-100 hospitals in digestive disorders, and 37 in heart and heart surgery respectively. As can be seen in Figure 1, in each of the three cases, the average quality score of hospitals where the chief executive officer is a physician is greater than the score of the hospitals where the CEO is a professional manager.

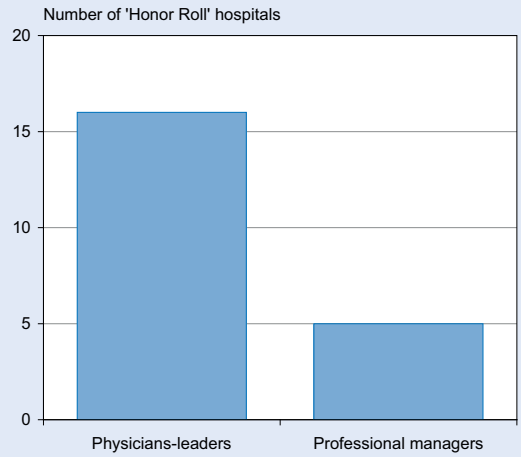
In the statistical analyses, the regression equations reveal that the presence of a physician-CEO is positively associated with an extra eight to nine hospital quality points (at the $p<0.001$ level) – in short, hospital quality scores are approximately 25 percent higher in physician-run hospitals than in the average hospital.

To control for the size of hospital, in the field of cancer I included a variable for the number of beds. However, this size variable was insignificant and, importantly, it did not affect the importance of physician-leaders.

The US News and World Report ranking also includes an 'Honor Roll' category which is made up of the most outstanding hospitals – those that achieved high hospital quality scores in at least 6 specialty fields. Figure 2 shows that the CEOs in 'Honor Roll' hospitals are more likely to be medically trained physician-leaders. Using a simple check I have found that in each year since 2009, when the data in this study were collected, 'Honor Roll' hospitals have continued to be dominated by physician CEOs.

Figure 2

Hospitals led by physician-leaders and professional managers in the elite 'Honor Roll'



Source: The author, based on US News and World Report ranking.

Why are better hospitals more likely to be led by physicians?

This study's results are cross-sectional associations and use one particular hospital-quality ranking. This means they have important limitations. The findings do not prove that doctors make more effective leaders than professional managers. Potentially, they may even reveal a form of the reverse – assortative matching – in that the top hospitals may be more likely to seek out MDs as leaders and vice versa. Arguably, however, the better hospitals will have a wider pool of CEO candidates to choose from thanks to the extra status and wealth that they attract. This makes the fact established in this study an interesting one. The results show that hospitals positioned highest in the ranking made judgements that differ from those made by hospitals lower down. On average they chose to hire physician-leaders as CEOs.These findings are consistent with my earlier work on the role of "expert leaders" in other (non-medical) settings as outlined above.

Cross-sectional analyses can only be suggestive of causality. It is nevertheless interesting to consider possible explanations. What differentiates expert leaders from generalists? Experts may have the advantage that they have acquired a deep intuitive knowledge about the core business of their organisa-

tions and this may help with decision-making and institutional strategy. Falcone and Satiani (2008, p. 92) suggest that a physician-leader who has spent years as a medical practitioner has acquired an integrity that implies "walking the walk" which, they argue, enhances a leader's credibility. Physician-leaders who have greater credibility may act as role models for medical staff and their presence may help hospitals to attract talented medical personnel. Hiring practices may be driven by homophily – like-for-like selection – thus, great surgeons and researchers may be more likely to hire other great surgeons and researchers. More importantly, it is probable that physician-leaders share the same values as other medically trained staff, and therefore may create better working conditions for doctors, surgeons and nurses.

There has been much journalistic coverage in the UK in recent years of the rise of managers and management practices in UK hospitals. UK hospitals are overwhelmingly led by non-MD managers. Might these manager-CEOs have been creating the right conditions for other managers, but not necessarily for their doctors? Such explanations are merely suggestive; as the mechanisms are not yet properly understood. The next, and vital, step for researchers is to design longitudinal inquiries into the possibility that physician-leaders improve the performance of hospitals.

Conclusion

There has been much discussion in the US, and increasingly in Europe, about the relative merits of having physicians and non-physician managers in leadership positions. Yet no evidence has been published one way or the other. This work does not establish that physicians make more effective leaders when compared with professional managers; but it starts the empirical process. It finds – in each of three disciplinary fields – that hospitals positioned higher in the US News and World Report's "Best Hospitals" ranking are led disproportionately by physicians. The next, and vital, step for researchers is to design longitudinal inquiries into the possibility that physician-leaders improve the performance of hospitals.

References

Candace, I. and R. W. Giordano (2009), "Doctors as Leaders", *British Medical Journal* 338, b1555.

Darzi, A. (2008), *High Quality Care for All: NHS Next Stage Review Final Report*, Department of Health, London.

Darzi, A. (2009), "A Time for Revolutions – The Role of Physicians in Health Care Reform", *New England Journal of Medicine* 361, e8.

Dwyer, A. J. (2010), "Medical Managers in Contemporary Healthcare Organisations: A Consideration of the Literature", *Australian Health Review* 34, 514–22.

Falcone, B. E. and B. Satiani (2008), "Physician as Hospital Chief Executive Officer", *Vascular and Endovascular Surgery* 42, 88–94.

Goodall, A. H. (2006), "Should Research Universities Be Led by Top Researchers, and Are They?", *Journal of Documentation* 62, 388–411.

Goodall, A. H. (2009a), "Highly Cited Leaders and the Performance of Research Universities", *Research Policy* 38, 1079–92.

Goodall, A. H. (2009b), *Socrates in the Boardroom: Why Research Universities Should Be Led by Top Scholars,* Princeton University Press.

Goodall, A. H. (2011), "Physician-leaders and Hospital Performance: Is there an Association?", *Social Science and Medicine* 73 (4), 535–39.

Goodall, A. H., L. M. Kahn and A. J. Oswald (2011), "Why Do Leaders Matter? A Study of Expert Knowledge in a Superstar Setting", *Journal of Economic Behaviour and Organization* 77, 265–84.

Goodall, A. H. and Pogrebna, G. (2012), "Expert Leaders in a Fast-moving Environment", *IZA Working Paper* no. 6715.

Gunderman, R. and S. L. Kanter (2009), "Educating Physicians to Lead Hospitals", *Academic Medicine* 84, 1348–51.

Horton, R. (2008), "The Darzi Vision: Quality, Engagement, and Professionalism", *The Lancet* 372, 3–4.

Pope, D. G. (2009), "Reacting to Rankings: Evidence from 'America's Best Hospitals'", *Journal of Health Economics* 28, 1154–65.

Schneider, E. C. and A. M. Epstein (1998), "Use of Public Performance Reports", *Journal of the American Medical Association* 279, 1638–42.

Smith, S. K. (2009), "The Value of Academic Health Science Centres for UK Medicine", *The Lancet* 373, 1056–58.

Stoller, J. K. (2013), "Commentary: Recommendations and Remaining Questions for Health Care Leadership Training Programs", *Academic Medicine: Journal of the Association of American Medical Colleges* 88 (1), 12–15.

Stoller, J. K., E. Berkowitz and P. L. Bailin (2007), "Physician Management and Leadership Education at the Cleveland Clinic Foundation: Program Impact and Experience over 14 Years", *Journal of Medical Practice Management* 22, 237–42.

US News and World Report (2009), *America's Best Hospitals Methodology and Ranking*, produced by Research Triangle Institute.

The Network Effect in International Migration

Michel Beine*

Introduction

Just like international trade and international capital flows, the international mobility of people is now part of the globalisation process. With downward pressure on tariffs and quotas and the implementation of trade agreements between countries, there has been an impressive increase in the international exchange of goods since the Second World War. The same trend has been seen in the international movement of capital for the last 30 years, triggered by the progressive eradication of various restrictions on capital mobility by most developed countries. For a long time, the international mobility of labour and people has been the missing link in globalisation. This has been identified as a major welfare loss by eminent economists like L. Pritchett and D. Rodrik from the Kennedy School at Harvard University. International migration has nevertheless experienced an unrecorded boom since the early 1990s. The total number of migrants between 1960 and 2010 has multiplied by roughly three, from about 77 million in 1960 to 214 million in 2010. Over 4.5 million people cross an international border to settle in a new country on an annual basis. A third of those migrants settle in an OECD country.

Stylised facts

Beyond the figures regarding the size of migration flows and stocks, there is also a clear trend towards an increase in the skill content of migrants. As reported by Docquier and Rapoport (2012), the number of highly educated migrants living in OECD member countries has increased by 70 percent since 1990, as opposed to 30 percent for low-skill migrants. The so-called South-North migration dynamic obviously dominates the global migration action, representing over 50 percent of all migration flows recorded at the world level (Özden et al. 2011). South-South migration involves more unskilled migrants and includes different types of agents like international refugees.

Questions

The above mentioned trends raise at least two important questions. Firstly, what explains the recent rise in the size of migration flows? Secondly, among the determinants of international migration flows, what are the most important factors shaping the skill content of international migrants? In order to address those questions, a screening of the extensive literature devoted to the determinants of international migration is necessary. This literature has for a long time uncovered the traditional key variables. These include the wage differential between the origin and the destination country, the bilateral distance defined in a broad sense (geodesic distance, common borders, language proximity, the existence of colonial links). A prominent role is also played by so-called pull and push factors. Push factors include origin specific developments that induce people to emigrate like climatic factors (the so called environmental migration), political instability and the quality of institutions or demographic factors. Pull factors include destination specific factors such as labour shortages and immigration policies.

One of the key questions is whether all of the above mentioned variables are able to account for a substantial part of the variability in the observed international migration flows? The answer is negative. One of the missing links is the network effect. The network effect might be defined as the global influence exerted by migrants at destination on the flows of newcomers from their origin country. A quick and simple example can easily illustrate the importance of the network effect. It also illustrates how the analysis might be flawed if this effect is not accounted for. In 1990 there were 194 Turkish migrants in Luxembourg, of which 44 percent had a tertiary edu-

* University of Luxembourg, IRES, CESifo and CREAM.

cation level. By contrast, there were 1,270,000 Turks in Germany, of which only five percent were highly educated, and 1,040 Turkish migrants in Spain, of which 33 percent could be considered highly educated. The interesting feature is that Turkey has no colonial link and no common language with any of these three destination countries. The immigration restrictions were and remain roughly the same, while the wage differential between Luxembourg and Germany is more or less equal too. How can such a gap in the size and the proportion of the population of skilled migrants in the two countries be explained? The answer is the Turkish diaspora in Germany, which generates some chain migration and explains the surge in migration flows between the two countries. It also partly explains why bilateral migration is dominated by unskilled migrants, unlike in Luxembourg and Spain.

Size and estimated elasticity

One important question is to what extent the influence of networks is significant on top of the role of the traditional factors mentioned above. To answer that question, two pieces of information are needed: figures relative to the size of migrants' networks and the value(s) of the elasticity related to the network effect. Firstly, the macroeconomic size of these networks is huge. Table 1, based on the Docquier, Lowell and Marfouk (2007) dataset on bilateral migration stocks by education level and updated with the 2005 data provided by the OECD, shows the figures for the main diaspora for the year 2000 and 2005. It also gives the proportion of skilled migrants (for 2000 only). These figures show that some diasporas like the Mexican diaspora in the US are really important. Furthermore, those figures tend to underestimate the true size of such diasporas for at least for two reasons: most figures only include legal migrants and permanent migrants. In some countries like Canada, temporary foreign worker programmes have expanded fast and the official figures may miss part of the action. Finally, some migrants like children under a certain age (often 15) are sometimes excluded from the figures. The most conservative estimates for the Mexican diaspora, for example, total around 14 million migrants, a twofold increase compared to the official figures for 2000. Table 1 provides the most important diaspora observed in 2000 and 2005 along with the proportion of highly-educated migrants in that diaspora (available only for 2000).

The second piece of information is provided by the empirical macroeconomic literature and takes the form of econometric estimates of the network effect. While there are obviously econometric challenges to be overcome in order to correctly estimate that effect, the few existing papers based on structural gravity models (Beine et al. 2011; Bertoli and Fernandez-Huerta Moraga 2012; Beine and Parsons 2012) come up with quite consensual estimates. At the global level (i.e. mixing up all types of flows) the elasticity is about 0.4. This means that, on average, a ten percent increase in the bilateral migration stock leads to a four percent increase in the bilateral migration flow over the next ten years. This elasticity jumps to 0.7 when we restrict our attention to migration to OECD countries (and to 0.9 if we restrict it further to the US as the migrants' destination). Breaking the figure down by skill level, the elasticity is about 0.6 for skilled migrants versus 0.8 for their unskilled counterparts. Furthermore, the share of variability in bilateral migration flows explained by networks at destination is quite important. By way of illustration, the share of explained

Table 1

Selected large diasporas (2005) and proportion of educated migrants (2000)

Origin	Destination	Size (2000)	Size (2005)	Proportion skilled (2000)
Mexico	US	6,374,825	10,668,900	14.4%
Turkey	Germany	1,272,000	1,568,700	4.8%
Philippine	United States	1,163,555	1,677,200	71.7%
United Kingdom	Australia	969,004	998,800	39.3%
China	United States	841,699	1,255,500	51.6%
India	United States	836,780	1,469,200	79.4%
Vietnam	United States	807,305	1,086,400	42.9%
Cuba	United States	803,500	946,500	38.3%
Canada	United States	715,825	907,900	61.4%
El Salvador	United States	619,685	1,032,700	18.3%
Algeria	France	512,778	1.305,900	10.2%

Sources: OECD (2012); Docquier et al. (2007, release 2.1).

variability by structural gravity models tends to fall by between 50 and 70 percent. At least one third of that proportion can be ascribed to the network effect, especially for unskilled migrants. This means that failure to account for the network effect in the modelling of the long-run mobility of workers results in a misspecified approach, and can lead to biased estimates of other determinants of migration.

Migrant networks as a selection device

The different elasticities across skill groups suggest that networks are not only an important determinant of the size of migration flows, but also act as a selection device in terms of the skill content of migrants. In other words, networks of migrants tend to reduce the proportion of skilled migrants in future migration flows. This has the opposite effect to other determinants like geodesic distance and selective migration policies. The existence of a strong network effect partly explains cases of so-called negative or intermediate selection in international migration. Negative selection of migrants tends to occur when migrants are less educated than natives in the origin country. Intermediate selection refers to cases in which migrants display more or less the same average skill level as natives. Selection nevertheless only refers to the first moment of the skill distribution. This does not imply similar distribution between migrants and natives. The degree of dispersion in the skill levels of migrants can be higher or lower depending on the specific migration process. Without networks, there is a clear trend towards the positive selection of migrants, as reflected by North-North migration. In order to further understand the reason for this, one needs to understand the main economic channels through which networks affect the migration process.

Channels

The network effect can be broken down into two main economic channels. The first one is called the assimilation channel and more or less covers the various ways in which people in a destination country can help newcomers. They can help new migrants to find an accommodation, comply with the legal constraints of the destination country and learn the local language. They can provide implicit insurance and give them informal jobs during hard times. There is also evidence of migrant clustering in formal jobs. Importantly, the magnitude of this assimilation channel significantly varies with education. It is much stronger for unskilled migrants, as shown by some microeconometric studies such as McKenzie and Rapoport (2010) in the case of Mexican migrants in the US. The second channel is immigration policy, and especially family reunification. In all developed countries, immigration policy gives new migrants the right to bring their relatives into the country. There is naturally a great deal of variation between types of migrants (temporary workers usually have limited rights), modalities (for example, the exact definition of relatives in the law) and destinations. Nevertheless, even in countries with explicit skill-biased immigration policies like Canada and Australia, the proportion of migrants arriving under kinship-based visas is not negligible. In 2010, family-based immigrants represented about 60 percent (resp. 58 percent) of the permanent immigrants in Canada (resp. Australia). Once again, this policy channel is stronger for unskilled migrants than skilled migrants. In a nutshell, highly-skilled migrants can easily migrate under an economic visa (H1B in the US for instance, through the point system in the UK) and do not need to rely on the family reunification scheme. For unskilled migrants from far away countries, a visa obtained through family ties is often the only alternative to illegal migration. All in all, these two channels explain why the network effect varies greatly across the skill levels of the prospective migrants.

Quantifying the relative importance of these channels is not an easy task given our poor measurement of immigration policies. Nevertheless, using an identification strategy based on the size of the various networks, Beine, Docquier and Özden (2012) show that the assimilation channel accounts for between 25 and 50 percent of the network effect. For unskilled migrants, this figure is close to 50 percent. For the US, there is also evidence that the importance of this policy channel has increased over time. This might be related to explicit changes in immigration policy, but also to episodes of legalisation of undocumented migrants.

Implications

The existence of a strong network effect has various important macroeconomic implications. Firstly and importantly, along with the presence of huge diaspora in a lot of countries, the existence of the network

effect implies a strong hysteresis in migration flows. The strong degree of chain migration means that the scope of action for migration policy in curbing some bilateral migration flows is rather limited. As an illustration of the strong dynamics imposed by networks, Table 2 presents a set of examples of pairs of countries for which both the bilateral migrant stock (in 2000) and the recent bilateral migration flow (in 2010) were the most important factors at destination.

Table 2

Examples of country pairs with largest migration stock and largest recent flow at destination

Origin	Destination	Diaspora (2000)	Annual flow (2010)-documented migrants only
Mexico	United States	8,250,000	139,120
Turkey	Germany	1,188,000	57,564
Algeria	France	1,210,600	19,135
Morocco	France	686,300	17,976
El Salvador	United States	750,000	18,806
Pakistan	United States	301,900	30,000
Tonga	New Zealand	165,00	751

Source: OECD (2012).

In some countries, there is an implicit or explicit objective of diversity across the origin countries of the migrants. Governments are often concerned with the excessive concentration of migrants from the same country. They fear the formation of migrant enclaves and suspect that huge diasporas slow the integration of migrants in the society. The network effect counteracts the integration objective and contributes to the concentration of new migrants in a limited set of important diasporas. In other words, while the network effect might increase the heterogeneity of the destination country's total population, it can also lower the ethnic diversity of migrants. In the same vein, a high concentration of migrants of the same country is observed in the big cities of destination countries. This is especially the case in large countries. Chinese migrants tend to concentrate in Vancouver, while Haitian ones mostly head for Montreal. Of course, policy reforms can be implemented to mitigate such an effect, but full eradication of family reunification rights is utopic. This means that one should not expect the concentration process to stop in the future.

A second implication is the impact of colonial links on current international migration flows. Unlike trade flows, colonial links have a rather indirect impact on contemporaneous migration flows. In the past colonial links made it possible to bring huge flows of people from the colonies, who settled permanently in the metropole after independence. Nowadays, new migrants from former colonies also tend to choose the former coloniser as their preferred destination, not because of previous colonial links (which often do not mean much to them), but because they receive support and are hosted by people of their origin country.

Sources of the network effect

Former colonial links are obviously one major source of the constitution of important diaspora in many destination countries. Algerians in France, Pakistanis and Indians in the UK, and Indonesian people in the Netherlands are perfect illustrations of the former colony phenomenon. However, colonial links are not the only source, as illustrated by huge networks like the Mexicans in the US, the Turks in Germany or the Portuguese in Luxembourg. A first alternative source is the past implementation of special bilateral agreements favouring worker's mobility between origin and destination countries. A perfect illustration is the broad category of guest worker programmes that were implemented in several European countries and the US after the Second World War to bring in workers in a set of specific industries suffering from labour shortages. The implementation of guest worker programmes were at the origin of diaspora like that of the Italians to Belgium or the Turks to Germany. Once again, when those programmes came to an end in the late 1960s and the early 1970s due to rising unemployment, those people had settled down and were already part of the population at destination. The existence of those guest worker programmes can be used as an exogenous source of variation of the network for the purpose of econometric identification and estimation of the network effect. This might be necessary because networks and current flows might be spuriously correlated due to their correlation with bilateral, persistent and unobserved factors such as cultural proximity. Another source of the huge diaspora lies in a perfect combination of skills at origin and needs at destination. Timing is also the key to gener-

ating such an effect. A good illustration is the Portuguese diaspora in Luxembourg. The boom in the construction sector in the late 1980s and in the 1990s in Luxembourg created a huge demand for those workers. A major part of that excess demand was satisfied by the arrival of Portuguese workers. This was also triggered by the detrimental business conditions in Portugal at that time, the relatively high reservoir of experienced construction workers and the fact that labour mobility was much easier between country members of the European Union. Today, the Portuguese diaspora in Luxembourg is by far the largest of its kind and represents about 16 percent of the Luxembourg's residents and 37 percent of all foreigners living in the country.

Implications for students and women

So far we have considered mainly economic migrants. Network effects are also relevant for subcategories of migrants such as students at the highest education level, as well as for women. It has been observed that foreign students of the same country tend to agglomerate not only in some specific destinations, but also in some universities. Quality of education, fees, language proximity and immigration policy all play important roles in that agglomeration process. However, networks are also part of the explanation. Networks operate at two different levels: firstly, student networks clearly provide useful information to newcomers regarding education programmes, education quality and future job prospects in the destination country. Secondly, diaspora can provide some useful hosting capacity in the form of accommodation. It is very valuable for students coming from developing countries with limited financial resources. For destination countries, this has important implications. In a globalising world there is sometimes fierce competition between countries to attract talents and skills, and attracting good foreign students is a successful strategy in this respect. Student migration is one indirect way to attract brainpower, with the additional advantage that the acquired skills are a better match to the needs of the local labour market. As far as women are concerned, new data on migration broken down by gender make it possible to characterise the migration processes involving men and women. Early studies showed that women are more sensitive to networks than men. This might, at first glance, be explained by biological differences. The common model is that of men taking foreign jobs and bringing their family with them afterwards. This is only part of the picture. Filipino nurses migrating in large numbers to the US and leaving children and husbands at home provide an important counter-example to that view (Filipino women represent about 60 percent of the Filipino migrants in the US). Secondly, different sensitivities to networks tend to disappear when they are made conditional to the education level of the migrants. In other words, skilled women and skilled men are equally sensitive to networks. One explanation is that women tend to be less educated than men on global average. While this is no longer the case in developed countries, it still applies in developing countries; and global migration is dominated by South–North flows, i.e. from developing to developed countries.

Limitations (and advantages) of macroeconomic approaches

So far, we have been concerned by the macroeconomic approach to the network effect in international migration. This is definitely not the only dimension and intellectual honesty leads the author to concede that this choice partly reflects some personal bias. Cross-country analyses deliver some clear advantages with respect to analyses focusing on single migration corridors. One of these advantages is that immigration policies can sometimes be accounted for explicitly. Moreover, the use of different origins and destinations makes it possible to increase the variability in some desirable dimensions such as education or gender. But cross-country macroeconomic analyses display obvious limitations that can be (partly) overcome by microeconomic approaches. Of course, as customary in the micro-macro debate, microeconomic data allow to control for the personal characteristics of agents. However, this is not the only key aspect here. Firstly, cross-country approaches implicitly assume that the relevant network is the total stock of migrants in the destination country. This is naturally an implausible assumption, especially in large destination countries. If you arrive in St Johns, New Foundland, Canada, it is very unlikely that your friends in Vancouver will be of valuable help. This implies that one needs to identify the size of the relevant network. Microeconomic data collected through surveys can be useful in that respect. The size of the relevant network operating through the assimilation channel may also differ depending on the exact type of effect that we are interested in. Assistance in providing accommodation is not simi-

lar to help in providing useful information. The use of microeconomic data makes it possible to reflect the topology of the network. The microeconomic literature of networks has expanded quickly during the last decade, both on the theoretical and empirical sides (see for instance Calvo-Armengol, Patacchini and Zenou 2009 on social networks and education outcomes and Zenou, 2012 for more general coverage of the literature on networks).

Bilateral links can be identified and can be used to measure the degree of connection of each individual in the network. This can be useful in estimating the relevant network elasticities in a more precise way. Furthermore, when properly collected, the use of microeconomic data makes it possible to circumvent tricky econometric challenges such as the reflection problem initially identified by Manski (1993).

Avenues of future research

In spite of a big surge in the number of economic analyses of the network effect in international migration, there is definitely some scope for further investigation in this area. The possibilities are very numerous and I will only focus on a couple. A first aspect that has been disregarded by the literature on this topic to date concerns the vintage issue of the network. Networks do not have the same age, and this affects their capacity to provide assistance to newcomers from their origin country. For the sake of illustration, the Italian diaspora in France, Belgium and Luxembourg is a relatively old one. Very often, inhabitants of Italian origin are fully assimilated in the population, often hold dual nationality and tend to have quite loose ties with their origin country. A significant number of these people hardly ever speak Italian and no longer have close family in Italy. This forms a stark contrast to the more recent Portuguese diaspora in Luxembourg. In this context, the network effect associated with those diaspora is likely to be different, both in terms of magnitude and in terms of the assimilation effects. The identification of the variability of those effects across different generations of network is a desirable avenue of research for the future. The identification of the peak in the time pattern of the network effect would be an interesting by-product of such an analysis.

Another avenue of research is the identification of global networks. Country-related definitions of networks can be too large, as mentioned above, but they can also be too narrow sometimes. People from different countries who speak the same language can provide some useful hosting capacity at destination. This is obvious in migration involving South American migrants. People from Ecuador can be of valuable help to newcomers from Columbia (and conversely, of course). The identification of the variables allowing for a more general definition of the relevant network is also a challenge for the next steps in the research in that field.

Last but not least, the microeconomic identification of key players in migration networks would also be an interesting avenue of research. Such research has recently been conducted in criminal networks and opens the door to further analysis in the field of international migration. The identification of the salient features of the agents playing an important role in the hosting of new migrants could definitely be of policy interest for governments. One policy implication of such a research agenda would be to identify the features that make a network successful (by helping new migrants, but also by favouring their integration within the destination country).

Conclusion

Academic research into the network effect in international migration has undergone major progress in recent times. This has been allowed by the creation of new data capturing the cross-country variation in bilateral migration stocks and flows. There have also been significant advances on the front of the relevant methodology to assess the importance of the network effect. Some valuable progress has been made in the development of micro-founded gravity models that allow for the identification of the theory-consistent determinants of the flows. Important concepts identified in the trade literature such as the multilateral resistance to migration have also been explicitly accounted for. However, this in no way rules out the need for further research. A first stone in the wall has been put in place in the form of consensual macroeconomic estimates of the network effect. These estimates need to be refined on several fronts, as proposed in the last part of this article.

References

Beine, M., F. Docquier and C. Özden (2011a), "Diasporas", *Journal of Development Economics* 95 (1), 30–41.

Beine, M., F. Docquier and C. Özden (2011b), "Dissecting Network Externalities in International Migration", *CESifo Working Paper* no. 3333.

Beine, M. and C. Parsons (2012), "Climatic Factors as Determinants of International Migration", *CESifo Working Paper* no. 3747.

Bertoli, S. and J. Fernandez-Huerta Moraga (2012), "Visa Policies, Networks and the Cliff at the Border", *IZA Discussion Paper* no. 7094.

Calvo-Armengol, A., E. Patachinni and Y. Zenou (2009), "Peer Effects and Social Networks in Education", *Review of Economic Studies* 76 (4), 1239–67.

Docquier, F., B.L. Lowell and A. Marfouk (2007), "A Gendered Assessment of Highly Skilled Emigration", *Population and Development Review* 35 (2), 297–321.

Docquier, F. and H. Rapoport (2012), "Globalization, Brain Drain and Development", *Journal of Economic Literature*, in press.

Manski, C.F. (1993), "Identification of Endogenous Social Effects: The Reflection Problem", *Review of Economic Studies* 60 (3), 531–42.

McKenzie, D. and H. Rapoport (2010), "Self-Selection Patterns in Mexico–US Migration: The Role of Migration Networks", *Review of Economics and Statistics* 92 (4), 811–21.

OECD (2012), *Resserrer les liens avec les diasporas: Panorama des compétences des migrants,* OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264178069-fr.

Özden, C., C. Parsons, M. Schiff and T. Walmsley (2011), "Where on Earth is Everybody? The Evolution of International Bilateral Migrant Stocks 1960-2000", *World Bank Economic Review,* 25 (1), 12–56

Zenou, Y. (2012), "Networks in Economics", *CEPR Discussion Paper* no. 9021.

Financing of the German Economy During the Financial Crisis[1]

Christa Hainz* and
Manuel Wiegand*

Introduction

After the break-out of the financial crisis the causes and the economic impact of credit constraints were widely discussed in the media and among academic researchers. In the narrowest sense, a firm is considered credit constrained when it is denied access to credit due to a supply-side shock from the bank-side, although the firm has profitable investment opportunities. Generally, the discussion centers on the question of whether firms, given that they had credit demand, had problems with access to credit during the financial crisis because of their own deteriorating creditworthiness (demand-side effects), or because of lending constraints at the bank-level (supply-side effects). Empirical research has shown that banks reduced the supply of credit to firms during the financial crisis (Ivashina and Scharfstein 2010; Jimenez et al. 2012; Popov and Udell 2012; deYoung et al. 2012). For Germany, the same has been found for retail lending by identifying banks that reduced lending because they themselves faced a liquidity shock (Puri, Rocholl and Steffen 2011).

As far as the effects of credit constraints are concerned, the question arose whether scarce bank credit is truly harmful or whether firms simply substitute bank credit with other financing instruments. Empirical research supports the hypothesis that restricted access to credit during banking crises has serious real economic effects (e.g. Reinhart and Rogoff 2009; Campello, Graham and Campbell 2010).

In this article we will provide a descriptive analysis of the credit financing impairments German firms faced due to the financial crisis, the importance of bank credit in the financing of firms and the role of their bank relationships. To assess these issues, the Ifo Institute conducted the "Financing of the German Economy" survey in September 2011. In the sample of 1,139 firms from the manufacturing sector that participated in the survey, small, medium-sized and large firms were evenly represented.

Credit financing impairments due to the financial crisis

To bring firms' perception of credit supply into the discussion about credit constraints during the financial crisis in Germany, firms in the Ifo "Financing of the German Economy" survey were asked whether they saw their credit financing impaired by the financial crisis. 22.1 percent of the firms surveyed confirmed that this was indeed the case. When it comes to assessing credit constraints, it is important to understand that the volume of bank credit granted in an economy can also decrease because firms have a lower demand for financing (for example due to business cycle fluctuations in the demand for its products or shocks from economic crises). Macroeconomic indicators like the volume of loans granted do not make it possible to disentangle supply and demand effects. Therefore, it helps to focus on firms that actually have demand for bank credit. In our sample many firms had not conducted loan negotiations since 2008. These firms were likely to report that they did not experience impairments caused by the financial crisis. Among the firms that negotiated a loan or a line of credit in 2008 or later, 31 percent reported credit financing impairments arising from the financial crisis.

If a firm reported that it experienced impaired credit financing, it was also asked what kinds of impairment it had faced. In the narrowest sense, credit constraints describe a situation whereby bank credit is not available to firms. As Figure 1 shows, over half of the firms with impaired credit financing in the Ifo "Financing of the German Economy" survey reported that the availability of new loans or lines of cred-

* Ifo Institute.
[1] We gratefully acknowledge financial support from the German Science Foundation (DFG) through project HA 3039/3-1.

it was an impairment caused by the financial crisis. In addition, 34.2 percent faced a reduction of existing credit lines, another indicator that the quantity of credit available was impaired by the crisis.

A second impairment due to the financial crisis was the increase in the interest rates charged for existing lines of credit or loans. This impairment was also reported by over 50 percent of the firms with impaired credit financing, which underlines that many firms still had access to bank credit, but they had to pay a higher price for it. This can firstly be explained by the fact that the business conducted by many firms became riskier during times of crisis. Banks used the higher interest rates to receive compensation for the risk incurred from lending to such firms. On the other hand, higher interest rates might also indicate that banks faced higher refinancing costs during the financial crisis and that these were passed on to their customers.

Besides the pecuniary transfer of interest rates, banks can require collateral if they lend to firms. The incentives to do so are twofold. Firstly, banks take over property rights of the collateral if a firm defaults on its debt and thereby limit their loss given default. Secondly, the prospect that a firm loses collateral if it does not service its debt provides an incentive for the firm to increase its effort to pay interest and repay credit. Therefore, collateral is a natural instrument for banks to control the risks of lending to firms during crisis times, and it is not surprising that 47 percent of the firms with impaired credit financing reported higher collateral requirements from banks.

To gain a deeper insight into collateral, firms in the Ifo "Financing of the German Economy" survey were also asked about the terms and conditions of the most recent loan or line of credit that they received. Figure 2 summarises what kinds of collateral firms had to pledge: 32.1 percent of all lines of credit and only 15.7 percent of all loans were granted without collateral. If collateral was pledged, the most prominent types of collateral were land and buildings. 41.9 percent of the lines of credit and 55.8 percent of the loans were collateralised with this kind of asset. In addition, other fixed assets are often used as collateral for loans, while lines of credit are more likely to be collateralised with current assets. While private property is rarely used, the results underline that guarantees are almost as important in credit financing among German firms as collateralising with current assets is.

Although constrained credit availability, higher interest rates and higher collateral requirements

Figure 1

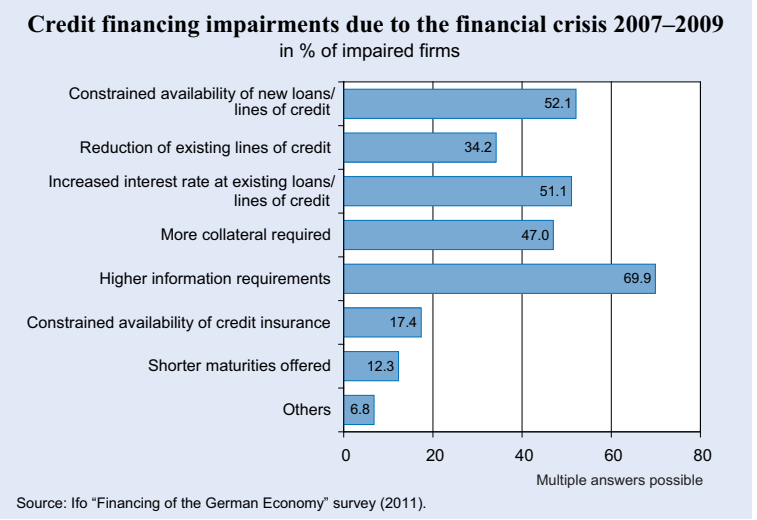
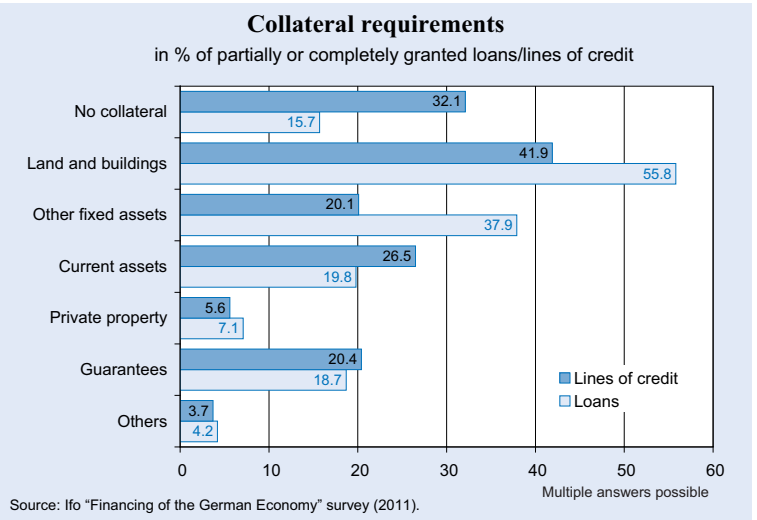


Figure 2



were an important impairment for the firms in the survey, Figure 1 shows that higher information requirements were reported by 69.9 percent of the firms with impaired credit financing. It is therefore the most frequent of all kinds of impairment. To resolve uncertainty about a firm's creditworthiness, banks require firms to provide information, for example through financial statements and business plans. During the financial crisis firms operated under high uncertainty, which made it even harder for banks to assess their creditworthiness. In response, many firms had to provide more information to banks, which created costs for the firms, as well as the banks, which had to process the information.

In addition to these frequent impairments, some firms also reported that they faced restricted availability of credit insurance (17.4 percent). Credit insurance is an important instrument for improving the availability of credit or improving the conditions whereby credit is granted. The impairment that banks only offered credit at shorter maturities was reported by only 12.3 percent of the firms with impaired credit financing.

The importance of credit financing

The negative impact of credit constraints on the German economy was widely discussed because bank credit is a major source of financing, particularly for small and medium-sized firms (SMEs). Since official statistics about the financing instruments used by German firms are scarce, the Ifo Institute asked firms which financing instruments they currently use. We find that 73 percent of all firms in the sample use bank credit. Almost half of the firms use leasing finance, about a quarter received loans from related firms and 10 percent use receivables financing (for example factoring). It is important to note that only 3.6 percent of the firms have access to capital markets (e.g. through the issuing of corporate bonds). Credit constraints may therefore have a serious effect on firms' investment and their business activity because their access to the capital market as an alternative source of funding in addition to bank credit is limited.

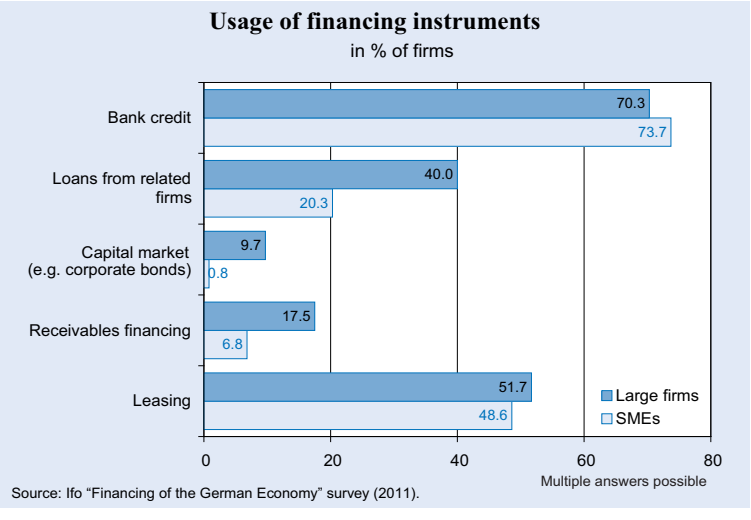
According to Figure 3, this is even more true of SMEs. For firms with less than 250 employees, bank credit is slightly more important than for large firms, but only 20.3 percent of SMEs use loans from related companies, 6.8 percent use receivables financing and only 0.8 percent have access to capital markets. These values are higher for large firms. Therefore, the financing structure underlines the importance of bank credit, in particular for SMEs, due to their limited access to alternative financing instruments.

When comparing this data to the financing structure of firms in other countries, one can see that the dependence on bank financing is a phenomenon that is not necessarily as present in other countries as it is in Germany, where a bank-based financial system is prevalent. For the US, as an example of a market-based financial system, a 2012 survey of the National Small Business Association has shown that only 43 percent of the firms used a line of credit to meet capital needs over the previous 12 months. Bank loans were even less common, with 29 percent of the firms using this instrument. Furthermore, it is interesting to note that credit cards are an important financing instrument for firms in the US (37 percent, see NSBA 2012).

The structure of the firms' bank relationships

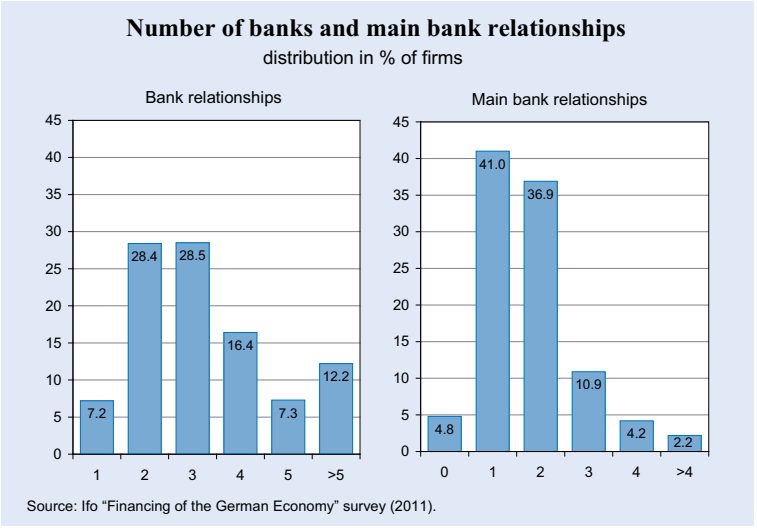
As bank credit is the most important source of financing for German firms, their decision about the

Figure 3



structure of bank relationships should also play an important role. Several empirical studies have analysed whether access to bank credit is affected by the number and the characteristics of bank relationships.[2] In the Ifo "Financing of the German Economy" survey a number of questions are devoted to the portfolio of bank relationships that a firm maintains.

Figure 4



In general, firms have to decide between focussing their business on a small number of banks to which close relationships are maintained and receiving financial products from a large number of banks without establishing close ties. On the one hand, a close relationship has the advantage that a relationship bank learns about the firm's creditworthiness over time, which may facilitate a firm's access to credit (Boot and Thakor 1994). On the other hand, if there is only one bank that knows the firm's creditworthiness, the firm depends on this bank because borrowing from uninformed non-relationship banks might be difficult (Sharpe 1990). The relationship bank can therefore develop an information monopoly. If a firm does not tie itself to a small number of banks, it forgoes the advantages of information provision within a close relationship, but can establish a better bargaining position against each bank because competition between banks is increased.

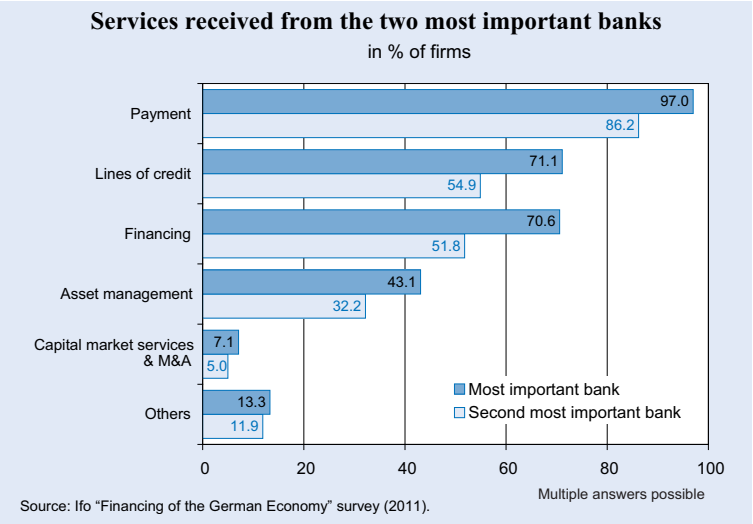
In Figure 4, the number of banks to which firms maintain business is summarised. The results show that only 7.2 percent focus all their business on one bank. The majority of the firms maintain two or three business relationships to banks. Almost half of the firms even have more than two banks.

To learn more about the character of these business relationships to banks, firms were asked how many banks they refer to as main banks (in German "*Hausbank*"). Close main bank relationships were raditionally perceived as an important characteristic of the German banking system. According to the survey, they are characterised by a long duration, personal support and a small distance between the bank and the headquarters of the firm. Only 4.8 percent of the firms do not have a main bank, which indicates that they do not follow the idea of focussing business on a relationship bank. The fact that over three quarters of all firms maintain one or two main bank relationships underlines the importance of focussing on a small number of very important banks among German firms. Referring to a larger number of banks as main banks, which would indicate a spread of business among many banks, is far less common. These numbers underline that firms in the sample tend to diversify their business relationships to banks, but at the same time maintain a small core of close long-term main bank relationships.

When asking firms about the characteristics of the two most important bank relationships (independent of whether the bank is a main bank or not), it becomes clear that the relationship between a firm and the key banks is stable over time. On average, the most important bank relationship has been established 29 years ago. The average length of the second most important bank relationship is 23 years. The survey also collected information on the products that a firm receives from the two most important banks (see Figure 5). The most important bank enjoys a leading position in the provision of all products. When looking at the differences, however, it seems that in particular core services like payment services, lines of credit and financing are more likely to be received from the most important bank than

[2] See Petersen and Rajan 1994; Berger and Udell 1995; Harhoff and Körting 1998; Cole 1998; Degryse and van Cayseele 2000; Lehmann and Neugerber 2001; Cole, Goldberg and White 2004; Santikian 2011; Bharath et al. 2011.

Figure 5



from the second most important institution. For capital market services, Mergers & Acquisition and other services the differences between the two banks are smaller.

In addition to these features of the two most important bank relationships, the data set contains information on the class of banks to which the two most important banks of every firm belong. This provides valuable information on how firms establish bank relationships within the institutional framework of the German banking system. Commercial banks, public banks and cooperative banks constitute the three different pillars of banks in Germany. Commercial banks are privately-owned universal banks that are to a large extent equity-financed. They operate without any regional restrictions and are often internationally active. The second pillar of the German banking system is the public banking sector. It consists of over 400 savings banks, each operating only within a certain region. They are owned by the respective municipalities and instead of profit-maximisation, their major goal is to take in deposits from local savers and lend to local borrowers. *Landesbanken* are also publicly owned and serve as central banks for the savings banks. They provide large-scale funding to private firms. The third pillar of the German banking system is the cooperative banking sector. A large number of small cooperative banks are only regionally active and their major goal is to serve their own members. Like savings banks, these cooperative banks focus on traditional banking activities. The DZ Bank and the WGZ Bank serve as central banks in the cooperative banking sector and offer financial services to firms that cannot be offered by small cooperative banks (for a more detailed description of the German banking system, see Hackethal 2004).

Table 1 shows how the choice of the most important bank depends on the size of the firm. For small firms savings banks are the most important class of banks followed by commercial banks and cooperative banks, which are equally important. *Landesbanken* and "Others" play a minor role. There are two potential explanations for the importance of savings banks and cooperative banks for small firms. Firstly, these firms might not need large scale funding so that savings banks and cooperative banks can provide credit to them without *Landesbanken* or commercial banks, which have the capacities to grant large-scale loans, needing to get involved. Secondly, if a firm is only active in a small local area, which is more likely for small firms than for larger ones, savings banks and cooperative banks with their dense branch network have a comparative advantage against large commercial banks when it comes to the assessment of the creditworthiness of firms within their region. Working with these local banks might therefore be advantageous to small firms.

Table 1

Classes of the most important bank of a firm (in % of firms)

	Employees			
	<50	50-249	>249	Total
Commercial bank	26.2	35.4	61.9	41.0
Savings bank	42.4	36.9	18.5	32.7
Cooperative bank	25.3	18.4	5.7	16.5
Landesbank	2.6	4.4	7.4	4.8
Others	3.5	4.9	6.5	5.0

Source: Ifo "Financing of the German Economy" survey (2011).

For medium-sized firms private banks become more important, mainly at the expense of savings banks and cooperative banks. This trend continues when looking at large firms for which the most important bank is a commercial bank in 61.9 percent of the cases, while savings banks are only half as important as for medium-sized firms and the importance of cooperative banks is negligible. Comparing all size groups, *Landesbanken* and "Others" are also most important for large firms. It is reasonable to assume that large firms need financing for larger projects, which might exceed the lending capacities of small savings banks and cooperative banks. Large firms also usually need a much broader range of financial services (for example capital market or foreign exchange products) that are not offered by savings banks and cooperative banks. In particular, when firms do business abroad they tend to establish relationships with banks that are internationally active as well. We can therefore conclude that the choice of the most important banks is clearly affected by a firm's size and the corresponding need for financial services.

Summary and conclusion

This descriptive analysis of the data from the Ifo "Financing of the German Economy" survey has shown that over 20 percent of the firms surveyed saw their credit financing impaired by the financial crisis, but impairments were experienced by 31 percent of firms with demand for new bank credit. Higher information requirements were the most frequent impairment, followed by constrained availability of new bank credit, higher interest rate payments and higher collateral requirements. It cannot be expected that these impairments were compensated by firms switching to other financing instruments because firms, in particular SMEs, reported that other financing instruments play a minor role in their portfolio. The survey also shows that although firms diversify their portfolio of business relationships to banks, they still tend to maintain a small number of main bank relationships to which close long-term relationships are established.

Further empirical analysis by Hainz and Wiegand (2013) shows that the focus on one main bank relationship helps to prevent some of the impairments listed above, namely higher information requirements, more collateral and shorter maturities, but not the others. In particular, the constrained availability of new bank credit and the reduction of existing lines of credit, which can be taken as symptoms of credit constraints in the narrowest definition, are not affected by a firm's focus on one main bank. These results stand in contrast to earlier work by Petersen and Rajan (1994), Harhoff and Koerting (1998), Cole (1998) and Cole et al. (2004), who find that a small number of bank relationships improves credit availability. These papers use data from the 1980s and 1990s. Hainz and Wiegand (2013) argue that the changes in lending technology and bank regulation that happened during the last 20 years limit the influence of soft information provided through a close bank relationship on the lending decision. If credit is granted, however, a close relationship can still be advantageous in the negotiation of the terms and conditions of the credit contract.

References

Bharath, S. T., S. Dahiya, A. Saunders and A. Srinivasan (2011), "Lending Relationships and Loan Contract Terms", *Review of Financial Studies* 24 (4), 1141–203.

Boot, A. W. and A. V. Thakor (1994), "Moral Hazard and Secured Lending in an Infinitely Repeated Credit Market Game", *International Economic Review* 35 (4), 899–920.

Campello, M., J. Graham and H. Campbell (2010), "The Real Effects of Financial Constraints: Evidence from a Financial Crisis", *Journal of Financial Economics* 97, 470–87.

Cole, R. A. (1998), "The Importance of Relationships to the Availability of Credit", *Journal of Banking and Finance* 22, 959–77.

Cole, R. A., L. G. Goldberg and L. J. White (2004), "Cookie Cutter vs. Character: The Micro Structure of Small Business Lending by Large and Small Banks" *Journal of Financial and Quantitative Analysis* 39 (2), 227–51.

Degryse, H. and P. van Cayseele (2000), "Relationship Lending within a Bank-Based System: Evidence from European Small Business Data", *Journal of Financial Intermediation* 9, 90–109.

deYoung, R., A. Gron, G. Torna and A. Winton (2012), "Risk Overhang and Loan Portfolio Decisions: Small Business Loan Supply Before and During the Financial Crisis", *Working Paper*, electronic copy available at: http://ssrn.com/abstract=2140952.

Hackethal, A. (2004), *German Banks and Banking Structure*, Oxford University Press, Ch. 3, pp. 71–105.

Hainz, C. and M. Wiegand (2013), "How does Relationship Banking Influence Credit Financing? Evidence from the Financial Crisis", *Ifo Working Paper*, forthcoming.

Harhoff, D. and T. Koerting (1998), "Lending Relationships in Germany - Empirical Evidence from Survey Data", *Journal of Banking and Finance* 22, 1317–53.

Ivashina, V. and D. Scharfstein (2010), "Bank Lending During the Financial Crisis of 2008", *Journal of Financial Economics* 97, 319–38.

Jimenez, G., S. Ongena, J.-L. Peydro and J. Saurina (2012) "Credit Supply versus Demand: Bank and Firm Balance-Sheet Channels in Good and Crisis Times", *European Banking Center Working Paper* no. 2012–003.

Lehmann, E. and D. Neuberger (2001), "Do Lending Relationships Matter? Evidence from Bank Survey Data in Germany", *Journal of Economic Behaviour & Organization* 45, 339–59.

NSBA (2012), Small Business Access to Capital Survey, available at http://www.nsba.biz/wp-content/uploads/2012/07/Access-to-Capital-Survey.pdf.

Petersen, M. A. and R. G. Rajan (1994) "The Benefits of Lending Relationships: Evidence from Small Business Data", *Journal of Finance* 49 (1), 3–37.

Popov, A. A. and G. F. Udell (2012) "Cross-Border Banking, Credit Access, and the Financial Crisis", *Journal of International Economics* 87 (1), 147–61.

Puri, M., J. Rocholl and S. Steffen (2011), „Global Retail Lending in the Aftermath of the US Financial Crisis: Distinguishing between Supply and Demand Effects", *Journal of Financial Economics 100* (3), 556–78.

Reinhart, C. M. and K. S. Rogoff (2009), "The Aftermath of Financial Crises", *American Economic Revie*w 99 (2), 466–72.

Santikian, L. (2011), "The Ties That Bind: Bank Relationships and Small Business Lending", *AFA 2011 Denver Meetings Paper,* electronic copy available at: http://ssrn.com/abstract=1718105.

Incentive Schemes for Local Government

Ben Lockwood* and
Francesco Porcelli*

Introduction

In recent years, explicit incentive schemes for public organisations, based on quantitative measurement of outputs, have become increasingly commonly used in the UK. For example, school league tables, hospital star ratings, and various schemes for local government, such as Comprehensive Performance Assessment (CPA), have been introduced in the last twenty years or so. Moreover, with few exceptions, schemes of this type have been little used outside the UK.[1] Finally, the schemes just noted have only been introduced in England, creating the possibility of using other regions of the UK as control groups to study their effects.

The focus of our work is on CPA, the most important scheme of this type for local government.[2] This scheme, introduced in 2001, rated local governments in England on the quality of service in six major areas: education, housing, social care, environment, libraries and leisure, and use of resources. Hundreds of performance indicators and a variety of audit and inspection reports were collected, summarised, weighted, and categorised so as to arrive at final star ratings between 0 and 4 stars.

As well as an evaluation scheme, CPA was also an incentive scheme. The stated objective of the CPA was to target support at those councils that need it most, and to offer a number of benefits for better-performing councils, including the elimination of "ring-fencing" grants, and a three-year exemption from subsequent audit inspections.[3]

Moreover, because the results of the CPA were widely disseminated in the media, it was also an exercise in providing voters with more information about the performance of their local council, both absolutely, and relative to other councils. In turn, this, in principle, provides indirect incentives for good performance. Indeed, there is evidence that councils which performed poorly on CPA were punished by voters at subsequent elections.[4]

CPA is of particular interest because it is, to our knowledge, the only explicit evaluation scheme to date, worldwide, that numerically scores and rewards elected representatives, as opposed to public service managers. The purpose of this paper is to assess the impact of CPA on local government in three dimensions: quality of service delivery, taxation policy, and the efficiency with which services were provided.

Figure 1 below shows the average CPA score achieved by English local authorities from the beginning to the end of the CPA experience together with average real current expenditure per capita by local government. There is clearly a steady upward trend in average CPA star ratings. Indeed, in 2009 the Audit Commission officially declared that the CPA had done its job stimulating a continuous improvement in local government performance (Audit Commission 2009). However, Figure 1 also shows that local government expenditure rose simultaneously, more or less in line with CPA scores.

* University of Warwick.

[1] There are exceptions: in the US, for example, the No Child Left Behind legislation punishes schools financially for poor test results, which are made public to parents.

[2] This report summarises findings from our paper, Lockwood and Porcelli (2013).

[3] "High scoring" councils were councils that were performing well under CPA and would consequently enjoy reduced audit and inspection regimes, and their associated fees, and be granted greater flexibility and borrowing freedoms by central government. At the other end of the performance spectrum, a combination of audit, inspection and other improvement work was to be commissioned as an outcome of the CPA process, with the aim of transforming failing or poorly performing authorities" (Audit Commission 2009).

[4] Revelli (2008) finds that an increase in one star rating increases the probability that the incumbent party retains control of the council by seven percentage points, and Boyne et al. (2009) find "a low CPA score (0 or 1 star) increases the likelihood of a change in political control".

So, the key problem is that we do not observe the counterfactual; given the large increases in local government spending over this period, it may be that service delivery would have improved anyway, even in the absence of the CPA. To address this, we treat the CPA as a natural experiment by exploiting the fact that it was only introduced in England, whereas in Wales, where the structure of local government is the same, a much weaker performance management scheme was introduced (Haubrich and McLean 2006b; Martin, Downe and Grace 2010). In particular, in Wales, there were no quantitative rankings, much less information published, and authorities also had a say with regard to the type of inspections they would like to see for specific services. So, we use local authorities in Wales as a control group when assessing the impact of CPA on the treatment group, the English councils.

What would we expect the effects of a scheme such as CPA to be on service quality, tax levels, and efficiency? In Lockwood and Porcelli (2013), we develop a simple two-period political agency model to focus specifically on the effect on taxation, spending and efficiency of an incentive scheme that both rewards service quality and provides information about this quality to voters. In any period, the quality of a public good or service is determined by a given politician's ability, efforts and tax revenue. In this environment, efficiency measures the level of service quality that can be produced at a given level of tax revenue. Voters value service quality and dislike taxes, and thus they care about both service quality and efficiency. The incumbent faces an election against a randomly selected challenger at the end of the first period.

Our key predictions (explained below in section *The effects of CPA – the theoretical predictions*) are as follows. The larger the direct reward, or the better the information provided by the incentive scheme, the more the incumbent politician taxes, and the higher the effort s/he makes. While greater effort is not surprising, the prediction of higher taxation, which voters dislike, is a distinctive feature of our theoretical analysis. As both effort and taxes rise, service quality is unambiguously increased by an incentive scheme. However, the effect of either a larger direct reward or better information on efficiency is ambiguous, because inputs, purchased using tax revenue, are also higher.

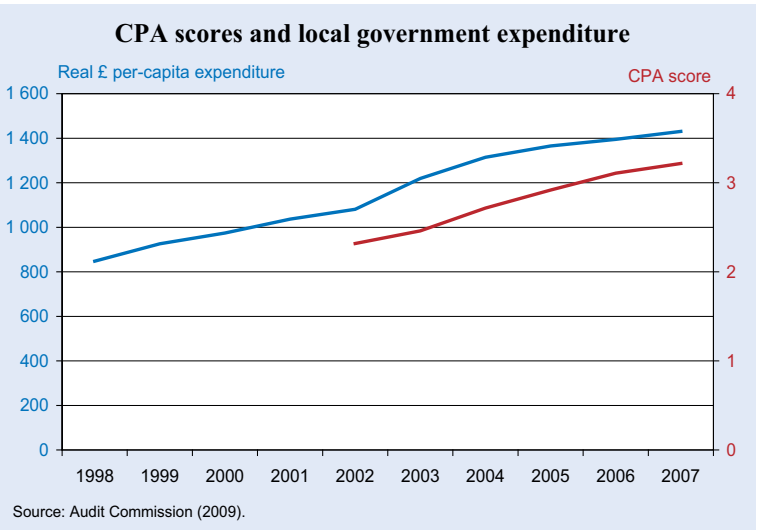
We then test these predictions using Wales as a control group. Our results broadly confirm the predictions of the theory, as described in more detail below.

The CPA – a brief overview

Local governments in England and Wales are of two types, unitary and two-tier. Unitary councils are responsible for primary and secondary education, social care, housing and housing benefit payments, waste disposal, transport, environment, planning, and culture. Two-tier governments are composed of an upper tier, counties, and a lower tier, districts. Counties have all the responsibilities of unitary authorities, except for housing and housing benefit, and environment, where responsibilities are shared with district councils.

In this institutional setting, the precursor to CPA, introduced in the Local Government Act 1999, was the "Best Value" framework, which, according to the UK government, "provides a framework for the planning, delivery and continuous improvement of local authority services. The overriding purpose is to establish a culture of good management in local government for the delivery of efficient, effective and economic services that meet the users'

Figure 1



Source: Audit Commission (2009).

needs."[5] A key part of this framework were the Best Value Performance Indicators (BVPIs), which were numerical scores measuring the quality of the above services on various dimensions. Importantly for our purposes, BVPIs were calculated for both English and Welsh councils.

CPA, which started in the 2001/02 financial year, represented a move to a stricter assessment regime within the general Best Value framework. In the first three rounds, the method for assessing the current performance of a council was as follows. Council performance was assessed in seven categories: social care; environment; libraries and leisure; use of resources; education; housing; housing benefit payments.[6] Where available, performance was assessed through already existing judgements from inspectorates and auditors, such as those by the Office for Standards in Education (Ofsted) and by the Department for Education and Skills (DfES) for education. These judgements were augmented with BVPIs. All this information was aggregated to obtain a score of between 1 and 4 for each of the service blocks (with 1 being the lowest and 4 the highest). The performance scores were then aggregated across service blocks to produce a performance rating of between 1 and 4 for each authority.[7] This score was then combined with an estimate of the councils' ability to improve (1 to 4) to produce the final CPA score.

In 2005, a new methodology, the "harder test", was introduced. The current performance of the council was assessed in the same categories with the exclusion of education, which was dropped. The main innovation, however, involved the aggregation procedure, where the ability to improve was replaced by the corporate assessment, a three year period assessment of the council's ability "to lead its local community having clearly identified its needs and set clear ambitions and priorities" (Audit Commission 2009).

So, what are CPA scores really measuring? Along with some commentators such as McLean, Haubrich and Gutiérrez-Romero (2007), we take the view that CPA is a hybrid measure, partly measuring levels of service quality (thorough the BVPIs), partly measuring operational efficiency (use of resources) and partly broader aspects of corporate health or effectiveness (ability to improve). In fact, Porcelli (2010) shows that councils' efficiency is only moderately correlated with CPA scores (a Spearman correlation of around 0.30), and inefficient local authorities can "buy" better CPA scores when favoured by a good local context.

Moreover, as McLean et al. (2007) point out, there may also be "categorisation errors" in the aggregation procedure in Table A3, where fine numerical scores are compressed into just four categories. So, we take the view that CPA scores measure both service levels (output) and efficiency, and do so with some error.[8] In this paper, we are not interested in CPA as a measurement system, but as an incentive scheme. That is why we construct our own, independent, measures of output and efficiency for local councils, with the aim of studying the effect of the CPA regime on those measures, along with taxation.

The effects of CPA

The theoretical predictions

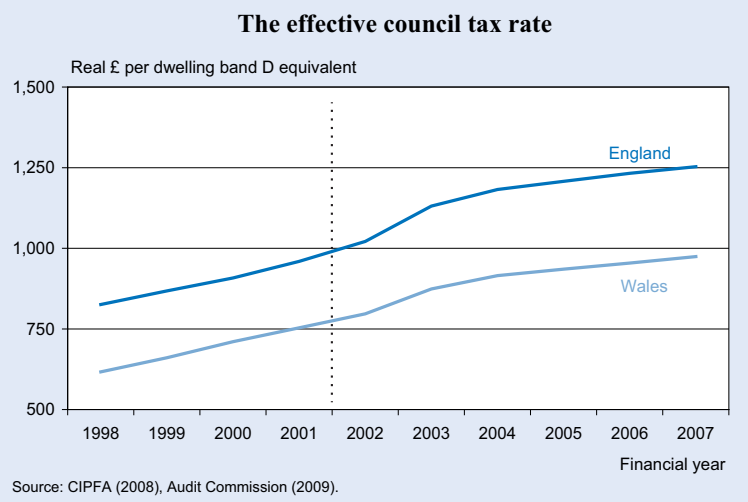
How might CPA be predicted to affect the behaviour of local governments in England? As discussed, CPA was a scheme that provided information to the voters (and also, possibly to the elected officials) of a jurisdiction about the quality and quantity of various "outputs" of local government. CPA may therefore be expected to cause these outputs to rise relative to those councils in Wales, our control group. However, funding from central government did not simultaneously become more generous in England relative to Wales. So the implication is that to fund this extra expenditure, taxes will rise in the "treatment group" i.e. in English local authorities. Finally, as argued above, CPA rewarded councils for overall increases in output, rather than increases in the efficiency with which inputs were used, so we should not expect to see any particular increase (or decrease) in the efficiency with which any council in England produces these services relative to a similar council in Wales.

[5] http://www.idea.gov.uk/idk.
[6] The CPA did not evaluate transport and planning.
[7] The scores were weighted so that the scores for education and social services count four times, housing and environmental services twice, with the remaining blocks counting only once. These were then added up to produce a performance score of between 15 and 60 points, or 12 and 48 points for shire county councils (because they do not provide, and are therefore not assessed on, housing or benefits services).

[8] Another possible source of error is that there is evidence that councils in areas where the population is more deprived or ethnically diverse achieve lower scores (Andrews 2004; Andrews et al. 2005; Gutierrez-Romero, Haubrich and McLean 2010): this may partly be due to higher (unobserved) costs of providing services in these environments.

with the same inputs and the same outputs for all units in all years, the only possible solution was to drop this sector from the efficiency analysis. A further problem is the short life of many BVPIs. Despite the fact that there are over 250 BVPIs published on the website of the Audit Commission, almost all of them were subject to some changes after three or four years, and in many cases they were replaced with new indicators. There is also the problem that after 2001–02, BVPIs were defined and measured separately in both England and Wales, and there was very little overlap. In the end, only five indicators could be used to measure the quality of output consistently for England and Wales; these measure aspects of education, social care of the elderly and children, waste disposal, and social services. However, it is important to note that expenditure on these categories accounts for fully 57 percent of total local government expenditure on average.

Figure 2



Source: CIPFA (2008), Audit Commission (2009).

Four of the five BVPIs are already expressed as percentages; and we also converted the fifth, social services to a percentage. We then calculated our output index as the weighted average of these five indices, where the weights used were the relative expenditure on the five services in real GBP per capita; all monetary amounts were deflated using the 2005 CPI. The source for the expenditure data is the Finance and General Statistics (FGS) and Local Government Comparative Statistics (LGCS), available on the website of the Chartered Institute of Public Finance and Accountancy (CIPFA) from the 1997/98 to the 2007/08 financial years (CIPFA 2008a and CIPFA 2008b).

Our efficiency index, denoted e_{it}, (where t refers to the time period, and i to the local council) is constructed using data envelopment analysis (DEA).[11] This method reduces the multiple inputs and outputs of any council in any given year to a single index. As output measures, we use the same five BVPIs used to construct the output index. As inputs, we use the expenditures already mentioned, corresponding to those outputs. Further details are given in our working paper, Lockwood and Porcelli (2011).

[11] DEA was first developed by Charnes, Cooper, and Rhodes (1978); a survey can be found in Ali and Seiford (1993).

DEA generates two indices. The first, the input index, e_{it}^{IN} , which lies between zero and one, has the following intuitive interpretation. If council i was using the available technology efficiently at time t, its inputs could all be scaled down by a fraction 1- e_{it}^{IN} and it would still be able to produce the same vector of outputs. The second, the output index, e_{it}^{OUT}, which also lies between zero and one, has a similar interpretation: if council i was using the technology efficiently at time t, its outputs could all be scaled up by an amount $1/e_{it}^{OUT}$-1, whilst using the same vector of inputs.

The input-based and output-based approaches to the evaluation of efficiency do not need to produce the same results; this will only occur in the restrictive case of constant returns to scale. Hence, in our analysis, the use of two indices can be considered as a sort of robustness check.

Empirical results

Taxes

Firstly, we look at the effect of CPA on increase council tax revenues. Figure 2 shows that the effective property tax rate (the tax requirement per standardised property) exhibits a clear increase in England relative to Wales after 2002. This is in line with what we would expect, based on our theoretical reasoning.

Of course, such a figure is only suggestive. A more formal analysis of the data is given in Table 1. The first two columns show the average values of the

Our full paper shows formally that the overall effect on efficiency is ambiguous, and identified conditions under which efficiency can increase or decrease.

Testing the theory using Wales as a control group

Our empirical approach is to estimate the impact of CPA on efficiency in a quasi-experimental setting through difference-in-difference estimation, using Wales, where CPA was not used, as a control group. Welsh local government performance was assessed by an evaluation program called the Welsh Program for Improvement (WPI) since 2001.[9] We believe that Welsh councils can be used to address the counterfactual question of what would have been the path of English councils after 2001 if CPA league tables had not been produced, for the following reasons:

Firstly, Welsh and English local authorities have the same structure and functions. Secondly, the average values of our control variables and the input and output variables used to construct our service quality and efficiency indices are very similar in the two countries. Thirdly, as documented by Haubrich and McLean (2006b), WPI was, compared to CPA, a much less prescriptive and elaborate assessment regime since only confidential assessments were produced, the evaluation criteria were based only on local self-assessment without quantitative rankings, and no formal rewards or punishments were specified. Finally, we have to address the question of whether the lack of "treatment" of Welsh local authorities was a truly exogenous event, or whether it was specifically related to the performance (in the setting of taxes or provision of public services) of Welsh councils. Firstly, the ability of Wales to determine a separate regulatory regime was ultimately determined by the creation of self-government in Wales, and in particular the creation of the Welsh National Assembly in 1998. Ultimately, support for devolution was determined by cultural factors, and can reasonably be regarded as exogenous. Secondly, as Haubrich and McClean (2006a) make clear, the main reason why the Welsh government did not adopt CPA was due to the smaller size of the country, which again is exogenous; "the relationship between auditor, local government department, and authority can be more intimate than in England".

Measuring tax revenue, output, and efficiency

Here, we discuss our choice of measures of taxes, output and efficiency for English and Welsh councils over our sample period 1997–2007. The data sources for these measures, and full details of how they were constructed, are to be found in our paper, Lockwood and Porcelli (2013).

The only tax instrument for local councils in the UK is a property tax; unlike in many other countries, there are no local income or sales taxes. The appropriate measure of tax is property tax revenue. This is measured by the tax requirement in the official statistics (CIPFA 2008a), which is total current spending in the financial year, minus revenue from the revenue support grant and other grants, and revenue from the business tax rate. We deflate this by the CPI to get real values.[10]

We use the tax requirement, both as a raw figure, and normalised in several ways. Specifically, we divide the tax requirement by the number of equivalent standardised properties (so-called "band D dwellings") to obtain an effective council tax rate. Finally, we also measure tax revenue as a percentage of the tax requirement to the budget requirement, where the latter is actual current expenditure that has to be financed by formula grants (which includes the police grant) and property tax revenue.

Next, we turn to the measurement of service quality. We need to construct a consistent index of service quality across both English and Welsh local governments. To this end, the BVPIs published by the Audit Commission for England and the Audit Office for Wales are the best source of information for two reasons: firstly they are broadly accepted by the local governments as measures of output quality; and secondly we are very confident about the comparability of these measures across local authorities since BVPIs were also chosen as one of the building blocks of the CPA procedure.

The first problem to solve was the absence of BVPIs for housing and housing benefit in case of the counties, where this function is managed by districts. As the efficiency analysis, further described below, analysis requires a balanced production function

[9] Information and data about the Welsh Program for Improvement can be accessed on the web site of the Wales Audit Office www.wao.gov.uk.

[10] Note that in England and Wales, local authorities can borrow only to finance capital spending, not current spending, and thus the difference between current spending and formula grants must be own revenues, principally the council tax.

effective property tax rate before and after the reform in both England and Wales. The third column shows the differences between the two, which are both positive. This is not surprising; we would expect taxes to rise over time, even in real terms. Finally, the last column shows that tax growth was significantly higher in England than in Wales during the period of CPA. In other words, there is evidence that CPA had asignificant positive impact on the effective property tax rate, raising it by an average of about GBP 52.

Of course, Table 1 reports just a simple difference in means, and there may be other factors driving relative changes in council taxes in England and Wales. In our full paper, we control for a large number of these factors. The first set of factors are demographic variables, such as the percentage of the total population below the age of 16 and above the age of 75, the percentage of population that declare itself religious, the percentage of white people, the population density, the percentage of households who own their house, and finally the tax base of the property tax (the number of band D equivalent dwellings per capita).

The second category includes a set of dummy variables to capture the impact of the ruling party and the features of the electoral system ("all out" election every four years, or "by thirds" system which involves more frequent elections). The third group of variables is related to the structure of the local economy and includes: the amount of real per-capita revenue support grant received every year by each council,[12] average household disposable income, the percentage of the workforce claiming unemployment-related benefits, the percentage of people below 65 claiming disability living allowance, the percentage of VAT tax payers in the financial and real estate sector, the percentage of highly-qualified workforce, and the percentage of the workforce that is self-employed.

We also control for business cycle effects or other unobserved time variation via year dummies. Finally, we consider the data as a panel i.e. we have four time observations before CPA, and six after, rather than just averaging observations before and after CPA.

After introducing these controls, we find that the effect of CPA on the council tax rate is slightly smaller, at GBP 46, corresponding roughly to a four percent increase in England relative to Wales. We also consider the effect of CPA on our two other measures of council tax revenues, the tax requirement per capita, and the tax requirement as a percentage of the budget requirement. The introduction of CPA raised the tax requirement by about GBP 23, or seven percent in England relative to Wales. Finally, it raised the tax requirement as a percentage of the budget requirement by about six percent in England relative to Wales.

Table 1

The effect of CPA on the effective rate of property tax

	Average pre-CPA	Average post-CPA	Difference	Difference-in-Difference
England	872.60	1,171.26	298.65	51.60***
Wales	662.15	909.20	247.05	

* significant at 10%, ** significant at 5%, *** significant at 1%

Source: The authors.

Outputs

We now turn to look at the effect of our output index, which is a variable normalised between 0 and 100, as described above. Figure 3 shows clearly that the output index rose faster in England than in Wales after the introduction of CPA.

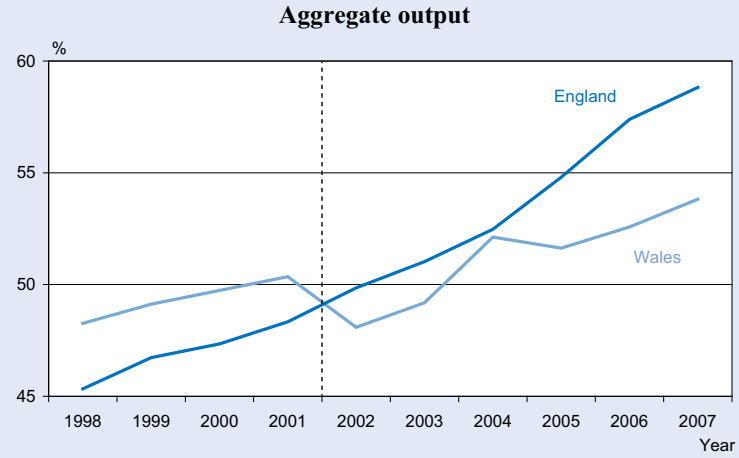
Again, we can investigate this further via a formal statistical analysis, which is presented in Table 2. The first two columns show the average values of the output index before and after the reform in both England and Wales. The third column shows the differences, which are both positive. That is, over time, councils in both England and Wales have managed to increase metrics such as exam performance, percent-

[12] It is important to stress that both the English and the Welsh grant system were based on the same rules during the period of our analysis. Differences only appeared in the English system after 2007. In particular, in both countries the system is formula based; grants can consequently be considered exogenous in relation to the behaviour of local governments, since they are mainly determined by local demographic and income characteristics.

age of waste recycled, etc. Finally, the last column shows that output growth was significantly higher in England than in Wales during the period of CPA. In other words, there is evidence that CPA had a significant positive impact on the output index, raising it by an average of about five percent

After introducing the large number of control variables already discussed, via multiple regression, we find that the effect of CPA on the output tax index rate is slightly smaller, at about four percent

Figure 3



Source: The authors.

Table 2

The effect of CPA on the output index

	Average pre-CPA	Average post-CPA	Difference	Difference-in-Difference
England	46.35	53.87	7.51	4.98***
Wales	48.85	51.39	2.53	

* significant at 10%, ** significant at 5%, *** significant at 1%
Source: The authors.

Efficiency

Let us now look at the effect of CPA on our efficiency indices. Figure 4 shows the path of the efficiency index in England and Wales (where the index is the average between the input and output approach) between 1997 and 2007. In both countries the initial decreasing trend in efficiency reversed its course after the introduction of CPA, and although the initial gap between Welsh and English councils is almost closed in the last year of the sample, there is no clear evidence that CPA has a positive impact on the efficiency of English local authorities.

Again, we can investigate this further via a formal statistical analysis, which is presented in Table 3 below. This analysis indicates two things: firstly,perhaps surprisingly, efficiency of provision of services has fallen over the CPA period in both England and Wales. Given that outputs have been rising, this implies that taxes and grants have been rising even faster. Secondly, there seems to have been no significant difference in the rate of change of the efficiency index in England and Wales.

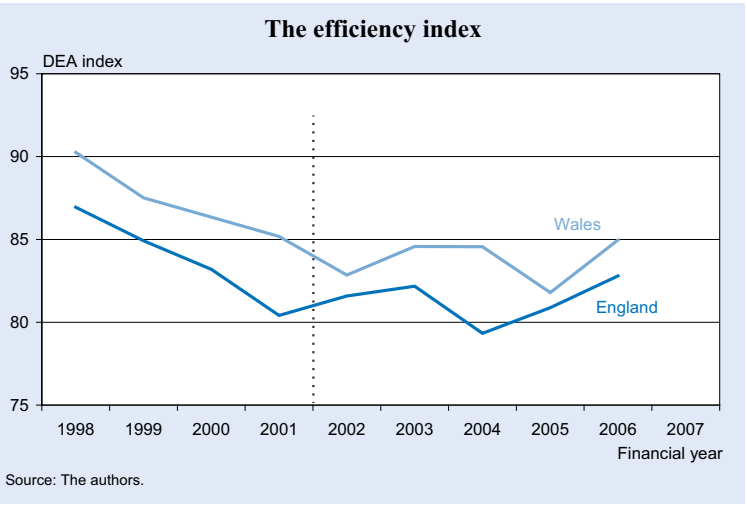
Robustness checks

A number of econometric robustness checks are reported in the paper. Here, we highlight two of these checks. One is to allow for council-specific time trends (see, for example, Friedberg 1998). To avoid collinearity problems, we add linear time trends for each type of council (London borough, Metropolitan district, County, Unitary authority, Welsh Unitary authority). The addition of these effects does not generally significantly change our regression results.

A second check, which is always important in a quasi-experimental setting, are placebo tests. Here, we run some placebo tests on the timing of the treatment. Specifically, we re-estimate the effect of CPA on output, tax and efficiency, assuming that the CPA program started in some other year than the year in which it actually occurred i.e. the fiscal year 2001/02. The results of these tests are also available on request, but we summarise them here. In the placebo treatments where CPA was introduced "before" 2001/02, either the treatment effect is insignificant or it has the opposite sign to that predicted by the theory i.e. negative effects on taxes and output. In the

placebo treatments where CPA was introduced "after" 2001/02, the treatment effect is mostly insignificant. However, we do observe significant positive treatment effects on taxes in cases where the placebo is one year after the true date of introduction. This could simply reflect the fact that councils reacted slowly to the introduction of the new regime.

Figure 4



Source: The authors.

Table 3

The effect of CPA on the efficiency index

	Average pre-CPA	Average post-CPA	Difference	Difference-in-Difference
England	84.41	81.33	-3.08	1.18
Wales	88.04	83.77	-4.26	

* significant at 10%, ** significant at 5%, *** significant at 1%
Source: The authors.

Electoral competition and CPA

The effects of electoral competition on policy-makers' behaviour are widely studied in the literature on political science, and increasingly also by economists. The study most closely related to ours in this respect is Besley and Preston (2007), who construct a measure of electoral districting bias for English local authorities. They find some evidence that a larger bias for the incumbent party (which protects the incumbent from electoral competition) gives the party a greater opportunity to pursue its policy preferences, which are lower expenditure and lower local government employment in the case of Conservatives, and the reverse in the case of Labour.

In our setting, it is plausible that CPA will have a larger effect on councils where electoral competition is low i.e. one party typically has a large majority of seats on the council. This is because such councils are initially not subject to much pressure to increase efficiency. So, in particular, we might find that efficiency is higher under CPA for low-competition English councils.

To test this, we define an English council to have "low electoral competition" if the winning party had a margin of victory over five percent. We can then Table 4 shows the change in the council tax rate, the output index, and the efficiency index over the CPA period (relative to the non-CPA period) for low-

Table 4

The effect of CPA on English councils with low electoral competition

	Change over CPA period		
	Council tax rate	Output	Efficiency
England (low competition)	286.70	8.30	-2.69
Wales	239.82	1.91	-4.17
Difference	46.87	6.38**	1.47***

* significant at 10%, ** significant at 5%, *** significant at 1%
Source: The authors.

competition English councils, and Welsh councils, the control group. The last line of the tables shows the differences between low-competition English councils, and Welsh councils in taxation, output, and efficiency.

Table 4 shows very clearly that low electoral competition has, in line with the theory, a significantly positive impact on both output and efficiency, but has no significant effect on tax. This result is robust to the inclusion of other control variables, and year dummies. However, there is no significant effect of CPA on tax levels.

So, the results indicate that CPA was a substitute for electoral competition; in councils where electoral competition was initially weak, it appears that CPA significantly increased both output and efficiency leaving the level of the property tax unchanged.

Conclusions

This paper has studied Comprehensive Performance Assessment, an explicit incentive scheme for local government in England, using Welsh local authorities as a control group, exploiting the fact that local authorities in Wales were not subject to the same CPA regime. We estimate that CPA increased the effective council tax rate in England relative to Wales by four percent, and also increased the index of service quality output by about four percent, but had no significant effect on our efficiency indices. Moreover, in line with the theory, there is robust evidence that CPA can substitute for an initial lack of electoral competition in driving up output and efficiency. The main policy implication of these results is that an incentive scheme like CPA can fail to stimulate higher local government efficiency because is too output-oriented; incentive schemes should be designed to place substantial weight on efficiency, and not just reward output.

References

Ali, A. and L. M. Seiford (1993), "The Mathematical Programming Approach to Efficiency Analysis", in A. O. Fried, A. K. Lovell, S. S. Schmidt, eds., *The Measurement of Productive Efficiency,* Oxford University Press, Ch. 3, 120–159.

Andrews, R. (2004), "Analysing Deprivation and Local Authority Performance: The Implications for CPA", *Public Money & Management*, 24 (1), 19–26.

Andrews, R., G. A. Boyne, J. Law and R. M. Walker (2005) "External Constraints on Local Service Standards: The Case of Comprehensive Performance Assessment in English Local Government", *Public Administration* 83(3),639–56.

Audit Commission (2009), Final Score: The Impact of CPA of Local Government, 2002-2008, www.audit-commission.gov.uk.

Besley, T. and I. Preston (2007), "Electoral Bias and Policy Choice: Theory and Evidence," *The Quarterly Journal of Economics* 122 (4), MIT Press, 1473–1510.

Charnes, A., W. W. Cooper and E. Rhodes (1978), "Measuring the Efficiency of Decision-Making Units", *European Journal of Operational Research* 2 (6), 429–44.

CIPFA (2008a), *Finance and General Statistics 2007/08*, The Chartered Institute of Public Finance and Accountancy, London.

CIPFA (2008b), *Local Government Comparative Statistics 2008*, The Chartered Institute of Public Finance and Accountancy, London.

Friedberg, L. (1998), "Did Unilateral Divorce Raise Divorce Rates? Evidence from Panel Data", *The American Economic Review* 88 (3), 608–27.

Gutierrez-Romero, R., D. Haubrich and I. McLean (2010), "To What Extent Does Deprivation Affect the Performance of English Local Authorities?", *International Review of Administrative Sciences* 76 (1), 137–70.

Haubrich, D. and I. McLean (2006a), "Assessing Public Service Performance in Local Authorities through CPA - A Research Note in Deprivation", *National Institute Economic Review* 197 (1), 93–105.

Haubrich, D. and I. McLean (2006b), "Evaluating the Performance of Local Government: A Comparison of the Assessment Regimes in England, Scotland and Wales", *Policy Studies* 27 (4), 271–93.

Lockwood, B. and F. Porcelli (2011), "Incentive Schemes for Local Government: Theory and Evidence from Comprehensive Performance Assessment in England", *The Warwick Economics Research Paper Series* 960, University of Warwick, Department of Economics.

Lockwood, B., and F. Porcelli (2013), "Incentive Schemes for Local Government: Theory and Evidence from Comprehensive Performance Assessment in England", *American Economic Journal: Economic Policy*, in press.

Martin, S., J. Downe and C. Grace (2010), "Validity, Utilization and Evidence-based Policy: The Development and Impact of Performance Improvement Regimes in Local Public Services", *Evaluation* 16 (31), 31–42.

McLean, I., D. Haubrich and R. Gutiérrez-Romero (2007), "The Perils and Pitfalls of Performance Measurement: The CPA Regime for Local Authorities in England", *Public Money and Management,* 111–17.

Porcelli, F. (2010), "Can Local Governments Buy a Good Performance Evaluation? Theory and Evidence from the Comprehensive Performance Assessment of English Local Authorities", unpublished paper, University of Warwick.

Revelli, F. (2008), "Performance Competition in Local Media Markets", *Journal of Public Economics* 92, 1585–94

The Bismarckian Factor: A Measure of Intra-generational Redistribution in International Pension Systems

Social systems, and especially pension systems, are commonly divided into two broad classes: they are organized according to the principles of either the *Beveridgean* or the *Bismarckian* tradition. Conceptionally, a Bismarckian pension system is characterized by a close link between previous earnings (and contributions when we assume that the latter are collected as payroll taxes) and today's benefits. A Beveridgean pension system, on the other hand, provides a basic or minimum pension. This binary characterization along the lines of the welfare state tradition ignores, however, that real-world pension systems typically contain elements of both. On the one hand, Bismarckian pension systems often provide benefits related not to previous contributions, but to personal characteristics (like motherhood, years at school) or earnings histories (like last or best contribution years). These exemptions tend to loosen the link between earnings and benefits, thereby inducing intra-generational redistribution. On the other hand, some countries with a strong degree of intra-generational redistribution have recently begun to introduce pension reforms that reduce this type of redistribution in their pension formulae. In order to evaluate how redistributive a pension system is, a more detailed measure of intra-generational redistribution is needed. The present data set aims to provide this information.

In order to determine the level of intra-generational redistribution in the public pension system ("first pillar"), we use micro data from the Luxembourg Income Study (LIS). The resulting measure of intra-generational redistribution may be referred to as the *Bismarckian factor*, following a convention in the theoretical contributions of, among others, Cremer and Pestieau (1998). Specifically, the index compares the inequality of pension benefits with the inequality of household net income, assuming that the principle of participation equivalence holds (see below). In a "pure" Beveridgean pension system, every pensioner receives the same pension benefit, independent of his/her (previous) household income. Here, the Bismarckian factor assumes a value of zero. Under a "pure" Bismarckian pension system, benefits are proportional to previous earnings/contributions, i.e., pension benefits exhibit the same level of inequality as earnings. Accordingly, the Bismarckian factor equals one.

Let Y^i and $P^i, i \in \{B(ottom),2,3,4,T(op)\}$, denote the mean income and the mean pension benefit, respectively, of the ith quintile of the income distribution. A purely Bismarckian pension system implies $\frac{P^B}{Y^B} = \frac{P^T}{Y^T}$, and a purely Beveridgean pension system implies $P^B = P^T$. The pension benefit of a representative member of quintile i, P^i, is defined as a convex combination of a flat payment (proportional to the mean income) and an earnings-related component (proportional to Y^i):

$$P^i \equiv \tau \cdot \left[\alpha Y^i + (1-\alpha)\mu\right], \qquad (1)$$

where $\alpha \in [0,1]$ is the Bismarckian factor, $\mu = \sum_i Y^i/5$ is the mean income of a society, and $\tau \equiv \sum_i P^i / \sum_i Y^i \in [0,1]$ the "generosity index", a measure of the level of redistribution between generations.

We are interested in a comparison of different income distributions at retirement age. Plugging equation (1) into the ratio of the pension benefits of the bottom and the top quintile, P^B/P^T, and solving for α gives:

$$\alpha = \frac{\mu(P^T - P^B)}{\mu(P^T - P^B) - P^T Y^B + P^B Y^T} \in [0,1]\,. \qquad (2)$$

A purely Beveridgean pension system yields $\alpha^{Bev} = 0$, while a purely Bismarckian pension system gives $\alpha^{Bis} = 1$. Hence, the Bismarckian factor is normalized on the closed interval [0,1]. Let us note that τ drops out of the formula, that is, the Bismarckian factor is independent of the generosity of the pension system. Accordingly, α is not only a pure measure of intra-generational redistribution but also allows for cross-country comparisons of public pension systems of different size. It is also worth noting that negative values of α can arise if pension benefits follow a progressive scheme due to, for example, means testing (formally, we then have: $P^B > P^T$).

All LIS data employed in computing the Bismarckian factor (Table 1)[1] and the generosity

[1] The tables can also be downloaded in the DICE Database under Social Policy / Pensions / System Characteristics.

Table 1

The Bismarckian factor

Country	LIS wave (years)							
	0 (-1978)	1 (1979-1983)	2 (1983-1987)	3 (1988-1992)	4 (1993-1997)	5 (1998-2002)	6 (2003-2004)	7 (2006-2008)
Australia	---	0.014	-0.086	0.046	0.113	0.010	0.029	---
Austria	---	---	---	---	0.501	0.525	---	---
Belgium	---	---	0.417	0.463	0.488	0.430	---	---
Canada	-0.002	0.035	0.046	0.066	0.270	0.307	0.289	0.265
Czech Republic	---	---	---	0.148	0.156	---	0.146	---
Denmark	---	---	---	---	0.056	0.024	-0.004	---
Finland	---	---	-0.044	0.019	0.594	0.416	0.364	---
France	---	0.710	0.701	0.711	0.730	0.737	0.715	---
Germany	0.573	0.579	0.583	0.539	0.564	0.589	0.549	0.575
Greece	---	---	---	---	---	---	---	0.638
Hungary	---	---	---	0.307	0.148	0.348	0.387	---
Ireland	---	---	0.121	---	0.347	0.327	0.348	---
Israel	---	-0.017	0.021	0.093	0.037	0.148	0.120	0.071
Italy	---	---	0.379	0.375	0.540	0.549	0.546	0.643
Luxembourg	---	---	0.445	0.367	0.315	0.351	---	---
Mexico	---	---	0.506	0.506	---	---	---	---
Netherlands	---	---	0.253	0.353	0.289	0.278	---	---
Norway	---	---	---	0.226	0.434	---	---	---
Poland	---	---	0.142	0.256	0.489	0.405	0.518	---
Slovenia	---	---	---	---	0.502	0.506	0.489	
Spain	---	---	---	0.528	0.432	0.470	---	0.554
Sweden	0.569	0.422	---	0.571	0.421	---	---	---
Switzerland	---	0.190	---	0.147	---	0.099	0.052	---
Taiwan	---	0.240	0.353	0.522	-0.068	-0.171	-0.152	---
United Kingdom	0.038	0.198	0.157	0.141	0.168	0.088	0.095	0.144
United States	0.340	0.342	0.532	0.533	0.545	0.462	0.445	0.461
Average	0.304	0.271	0.283	0.329	0.351	0.328	0.290	0.419

Note: A dash means that no data set (or first-pillar pension data) is available for the respective LIS wave.
LIS = Luxembourg Income Study.
Source: The authors.

index (Table 2) refer to the household level. We use "raw" household net income. Hence, α and τ measure the *legal status* of the pension system as it is reflected in the respective income distribution. This means that the numbers reported in the tables do not account for differences in needs due to household composition. Let us also note that we use the household weights provided by LIS in order to weight cases, if available. LIS reports household net income in an aggregate variable (DPI). The first-pillar of the pension system (i.e., the public part) is captured by three variables: HMITSILEP contains employment-related old-age, disability, and survivors' public pensions; HMITSUP and HMITSAP contain the respective figures for non-employment-related public pensions (universal pensions and social assistance, respectively).

LIS data is organized in "waves", that is, a data set is assigned to a certain wave if its base year falls into the respective time period, which usually comprises five years. Sometimes, and for some countries, several data sets are available referring to the same wave. In such cases, we selected one data set according to two criteria: firstly, only data sets for which the relevant pension variables were available were considered and, secondly, among those data sets we chose the eldest. As shown in the tables, there are many waves for which data sets and/or the respective variables are not available for some countries. Since LIS data rely on different samples for each wave (cross section), we cannot directly compare a single individual's pension benefit with his/her previous earnings, but have to resort to income distributions instead. This implies that our estimates of the

Table 2

The generosity index

Country	LIS wave (years)							
	0 (-1978)	1 (1979-1983)	2 (1983-1987)	3 (1988-1992)	4 (1993-1997)	5 (1998-2002)	6 (2003-2004)	7 (2006-2008)
Australia	---	0.065	0.080	0.054	0.071	0.068	0.067	---
Austria	---	---	---	---	0.209	0.224	---	---
Belgium	---	---	0.172	0.180	0.217	0.198	---	---
Canada	0.042	0.051	0.066	0.076	0.041	0.044	0.045	0.044
Czech Republic	---	---	---	0.206	0.182	---	0.214	---
Denmark	---	---	---	---	0.185	0.174	0.175	---
Finland	---	---	0.066	0.058	0.262	0.224	0.224	---
France	---	0.171	0.198	0.209	0.240	0.233	0.213	---
Germany	0.199	0.193	0.188	0.207	0.202	0.209	0.210	0.204
Greece	---	---	---	---	---	---	---	0.254
Hungary	---	---	---	0.220	0.040	0.280	0.280	---
Ireland	---	---	0.057	---	0.119	0.115	0.118	---
Israel	---	0.054	0.075	0.059	0.062	0.077	0.079	0.073
Italy	---	---	0.209	0.216	0.251	0.265	0.267	0.296
Luxembourg	---	---	0.223	0.210	0.216	0.206	---	---
Mexico	---	---	0.022	0.024	---	---	---	---
Netherlands	---	---	0.165	0.166	0.175	0.155	---	---
Norway	---	---	---	0.163	0.132	---	---	---
Poland	---	---	0.139	0.169	0.305	0.330	0.345	---
Slovenia	---	---	---	---	0.246	0.249	0.253	---
Spain	---	---	---	0.200	0.204	0.201	---	0.185
Sweden	0.107	0.252	0.272	0.256	0.215	---	---	---
Switzerland	---	0.084	---	0.112	---	0.160	0.122	---
Taiwan	---	0.001	0.001	0.005	0.004	0.007	0.011	---
United Kingdom	0.087	0.119	0.111	0.100	0.110	0.091	0.091	0.085
United States	0.076	0.054	0.101	0.102	0.113	0.098	0.104	0.104
Average	0.102	0.104	0.126	0.142	0.165	0.172	0.166	0.156

Note: A dash means that no data set (or first-pillar pension data) is available for the respective LIS wave.
LIS = Luxembourg Income Study.
Source: The authors.

Bismarckian factor should be interpreted according to the principle of participation equivalence. This principle assumes that intra-generational redistribution takes place if the individual replacement ratio (defined within the same wave) of a pensioner decreases with her individual benefit. This requires the (weak) assumption that a complete re-ranking of income and benefit positions will not take place between any two consecutive generations. Based on this assumption, it is justified that our data set rests on a comparison of today's earnings and today's pension benefits. This makes our data set especially useful for the analysis of distributional conflict since the participation of a pensioner in today's societal activities depends mainly on his/her personal position in terms of income distribution and the relative income position of pensioners compared to income earners.

Further information on the Bismarckian factor and an empirical application can be found in Krieger and Traub (2011). Please note, however, that the figures reported there do not accord with Tables 1 and 2 due to major data and variable revisions recently carried out by LIS.

Tim Krieger (University of Freiburg) and
Stefan Traub (University of Bremen).

References

Cremer, H. and P. Pestieau (1998), "Social Insurance, Majority Voting and Labor Mobility", *Journal of Public Economics* 68, 397–420.

Krieger, T. and S. Traub (2011), "Wie hat sich die intra-generationale Umverteilung in der staatlichen Säule des Rentensystems verändert? Ein internationaler Vergleich auf Basis von LIS-Daten" *Jahrbücher für Nationalökonomie und Statistik* 231/2, 266–87.

Unit Labour Costs in the Eurozone

One reason for the on-going euro crisis is the difference in the development of the average cost of labour per unit of output and its effect on international competitiveness. The OECD derives those average costs from Unit Labour Costs (ULCs), which are calculated as the ratio of total labour costs to real output (OECD 2013). This ratio can also be described as the ratio of mean labour costs to labour productivity, so that ULCs link productivity to labour costs. This link makes ULCs crucial for international competitiveness. The following text examines differences in the development of ULCs in member states of the Eurozone[1] over the last twenty years based on OECD data. The development of UCLs is discussed at three different time periods: in the 1990s, after the introduction of the Euro and at the beginning of the crisis. The main focus will be on trends in ULCs in the crisis-hit countries (Greece, Spain, Portugal, Italy and Ireland), as well as on their development in Germany. The base year for all data is 2000.

From 1992 to 2000, the development of ULCs in the crisis-hit countries was already stronger than in other member states. Greece and Spain in particular, with respective levels of 59.9 and 82.2 percent of ULCs 2000 in 1993, experienced higher ULC growth rates in the 1990s than other states of the subsequently formed Eurozone, where rates ranged between 89.3 percent (Luxembourg) and 96.0 percent (Finland). Even at that time Germany's ULC development was outstanding. Its ULCs of 1993 were just 2.1 percent below the level reached in 2000, meaning that costs more or less stabilised over this period. The only country in which a similar trend was observed was Austria.

The introduction of the Euro as a common currency in 1999 led to a credit boom in the crisis-hit countries, caused by fast growth and higher public spending in a context of higher inflation compared to the Eurozone average. The inflation led to a loss of international competitiveness, as shown in the OECD statistics, due to higher prices accompanied by higher costs and the lack of any possibility of debasement. In the second observed period from 2000 to 2007, which ends on the eve of the crisis, the ULCs of Greece (128.6 percent), Spain (127.7 percent), Italy (123.6 percent), Portugal (122.3 percent) and Ireland (133.6 percent) rose strikingly more sharply than in other states, where the increase generally totalled between five and 15 percent. Germany again stood out as having almost constant ULCs from 2000 to 2003 and as seeing a decrease in ULCs, which fell below the level of 2000 by 2007. This stand-alone development is the result of labour market reforms undertaken in Germany in the context of the Agenda policy pursued during Schröder's second term in office.

In the third period since 2007, ULCs rose in the five crisis-hit states to a level ranging from 130.5 percent (Portugal) to 144.1 percent (Greece). Since 2009 the trend has changed and ULCs have started to fall[2], bringing ULCs in 2012 back to the level of 2007[3]. The only exception to this rule is Italy, where ULCs rose continuously. The decline of ULCs in crisis-hit countries has been accompanied by an increase of ULCs in Eurozone countries since 2007. Except for the period from 2009 to 2010, during which a little decrease based on the crisis can be observed, the trend of rising ULCs has continued to date. Even in Germany, ULCs rose to 108.0 percent of 2000 for the first time in 2012. While the decrease in the costs of the crisis-hit countries depends heavily on the reforms implemented, the increase in wages, especially in Germany, is based upon the positive economic climate of recent years. This slower growth in ULCs led to a reduction in trade balance deficits in recent years, due to decreasing export prices associated with a decrease in labour costs.

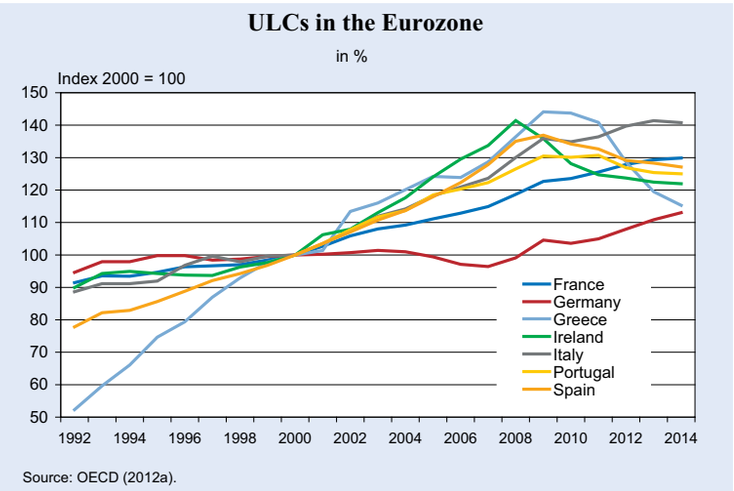
According to OECD predictions (see Figure 1), this downside trend in ULCs will continue for the next two years, so that Spain (127.0 percent), Portugal (125.0 percent), Ireland (121.9 percent) and Greece (115.0 percent) will – compared to 2000 – have smaller ULCs growth rates than most other countries, which will have a level of over 130 percent in 2014. The only exceptions are Austria (122.9 percent) and Germany, which with a level of 113.0 percent are again clearly below the average.

Since the prior adjustment of the trade balance deficits in the crisis-hit countries was attenuated by weak domestic demand and high unemployment, the

[1] I only look at the countries that introduced the Euro in 1999.

[2] In Ireland, the ULCs already started to fall in 2008.
[3] In Portugal, they are on the level of 2008.

Figure 1



Source: OECD (2012a).

OECD sees further need for structural reforms in the crisis-hit states. The OECD recommends economically stronger countries (OECD 2012b) – especially Germany – to implement further reforms with a policy featuring wage adjustment and a policy that stimulates demand in these countries. If further developments occur as predicted, the pressure on the crisis-hit countries exerted by ULCs will decrease.

Martin Voggenauer

References

OECD (2013), Main Economic Indicators: Sources and Definitions, http://stats.oecd.org/mei/default.asp?lang=e&subject=19.

OECD (2012a), Economic Outlook No 92 - December 2012 - OECD Annual Projections, http://stats.oecd.org/index.aspx?DataSetCode=EO92_INTERNET.

OECD (2012b), Euro Area Labour Costs Converging, but Imbalances Persist, http://www.oecd.org/economy/euroarealabourcosts.htm.

The Gender Wage Gap in OECD Countries

On average, women still earn considerably lower wages than men in all OECD countries. This is confirmed by the recent OECD Employment Outlook 2012. A measure of unequal pay for women compared to men is the so-called gender pay gap. It is reported as the difference in the median earnings of men and women relative to the median earnings of men. Table 1 presents the gender wage gaps for selected years and OECD countries. The numbers refer to the gross earnings of full-time employees.

In most countries the wage gap was between ten percent and 20 percent in 2010. Japan, however, boasted the largest gap of 28.7 percent, followed by Germany with 20.8 percent. In the US and the UK wage gaps are also quite high, at 18.8 percent and 18.4 percent respectively. Earnings differentials are less pronounced in some Eastern European countries like Hungary (6.4 percent) and Poland (6.2 percent). A low wage differential of 6.1 percent was also reported for Spain. Scandinavian countries, like Denmark (11.8 percent) and Norway (8.1 percent) boasted relatively small wage gaps too.

In all OECD countries, however, women earn persistently less than men when median full-time wages are used as a basis for comparison. Over time, earnings differentials have only decreased slightly. The greatest improvement can be seen in Spain, Ireland and the Netherlands. The OECD mentions that, in general, the differences in pay are less pronounced for new labour market entrants than for older age groups. The wage gap is highest among the tertiary educated.

Legislation to ensure equal pay for equal work regardless of gender that has been implemented in almost all OECD countries has not yet been sufficient to close the wage gap. Hence, what are the reasons for such persistent earnings differentials between men and women?

According to human capital theory, differences in pay can be explained by differences in individual characteristics like education, experience and age. However, evidence suggests that these factors alone do not shed much light on the earnings differentials observed. For example, the average levels of education of men and women in OECD countries are quite similar (OECD 2012b). There are considerable differences, however, when looking at the fields of education and the occupations chosen by men and women. Several occupational fields and industries are typically dominated by men, and others by women (OECD 2012b).

According to a review by the European Commission (2006) this labour market segregation is a major reason for the gender wage gap: relatively more women self-select into jobs with lower wages. Economic theory would suggest that earnings differentials due to labour market segregation reflect productivity differences between occupations and industries in the economy. Accordingly, many regression analyses explaining the gender wage gap control for occupations and industries, as well as for human capital explanatory factors like education and experience. The residual part of the wage gap only explained by gender differences would then reflect direct wage discrimination against women. The OECD Employment Outlook 2008 finds that, on average, around 70 percent of the unadjusted wage gaps in OECD countries can be explained in such a regression framework by non-gender related variables. However, 30 percent of the wage differentials are estimated to be due to discriminating practices against women on the labour market.

Other research[1], however, already considers labour market segregation and gender specific differences in experience and education as a form of discrimination. Arguments supporting this view refer to different wage structures in female and male dominated industries, gender specific education incentives and lower labour market experience of women due to childcare, for example. Actual wage discrimination against women would then account for more than the estimated 30 percent of the unadjusted gender wage gap.

Thus, only part of the wage gap can be explained by individual, labour market related characteristics. Direct and indirect wage discrimination, on the other hand, seems to have a considerable impact on female earnings.

Various policy measures already target towards closing the gender pay gap in OECD countries and

[1] For and overview on different publications see European Commission (2006).

reflect the complexity of the problem. Several approaches aim at enhancing equal opportunities and more continuous employment patterns for women.

Proving public day care possibilities, for example, is important to allow mothers to participate in the labour market. Moreover, some parental leave policies target more equal division of childcare between men and women. Denmark implemented such a policy in 2002. Some OECD countries, on the other hand, still lack such family policies. The large gender pay gap in Japan, for example, is mainly due to high wage penalties for mothers (OECD 2012a). A similar situation is reported for the Korean labour market.

Policy-makers in some OECD countries, like Germany, discuss quota systems to support women's access to managerial positions and, thus, reduce the gender pay gap. Norway, for example, has implemented a quota system since 2003 for boards of management in large, stock market listed companies.

A detailed table on policies addressing the gender pay gap in OECD countries can be downloaded from the DICE online database.[2]

As Table 1 shows, the gender wage gap has mostly decreased, but is still pronounced in many OECD countries. Earnings differentials that are due to unequal opportunities and wage penalties for women have to be further addressed by policy makers. Improving the earnings situation and incentives for labour market participation of women is even more important in the light of the demographic challenges in many OECD countries. Those economies cannot afford to lose human capital and economic potential. This is even more the case as highly educated women are most affected by the gender wage gap. Both governmental policies and social partner actions are necessary to overcome the challenges mentioned above. Beyond government policies wage setting institutions have to be involved in particular when addressing wage differentials between gender segregated labour markets.

Till Nikolka

Table 1

Gender wage gaps in OECD countries

	Gender wage gap (%)					
	1996	1998	2000	2006	2008	2010
Austria	35.0	23.0	23.1	22.0	21.0	19.2
Belgium		15.0	13.6		10.0	8.9
Czech Republic		25.0	21.8		21.0	18.1
Denmark	14.0	15.0	14.7	11.0	12.0	11.8
Finland	20.0	21.0	20.4	19.0	21.0	18.9
France		9.0	9.5		12.0	14.3
Germany	24.0	22.0	21.0		25.0	20.8
Greece					10.0	12.2
Hungary		16.0	14.1		2.0	6.4
Ireland	22.0	22.0	19.7	14.0	* 16.0	10.7
Italy			7.4		1.0	10.6
Netherlands	22.0	22.0	21.4	19.0	17.0	16.7
Poland	17.0			11.0	14.0	6.2
Portugal					16.0	13.5
Slovak Republic						14.8
Spain					12.0	6.1
Sweden	16.0	17.0	15.5	15.0	15.0	14.3
United Kingdom		26.0	25.5	21.0	21.0	18.4
Norway		10.0	10.2		9.0	8.1
Switzerland	25.0	22.0	22.2		20.0	18.5
Australia	15.0	13.0	17.2	17.0	12.0	14.0
Canada	25.0	25.0	23.9	21.0	20.0	18.8
Japan		35.0	33.9		31.0	28.7
New Zealand		11.0	7.1		8.0	6.8
United States	25.0	24.0	23.1	19.0	20.0	18.8

Empty cells: Data not available. *Preliminary estimate.
Estimates of earnings used in the calculations refer to gross earnings of full-time wage and salary workers.
The gender wage gap is calculated as the difference between median earnings of men and women relative to median earnings of men. Data refer to 1997 (instead of 1996) for Australia, Canada and Ireland, to 1998 for Poland; also to 1997 (instead of 1998) for Ireland, to 1999 for Belgium and to 2000 for Austria. They refer to 2002 (instead of 2006) for the Netherlands; to 2004 for Poland and Sweden; to 2004 for Finland, to 2005 (instead of 2008) for the Netherlands and to 2007 for Belgium and France. They refer to 2001 (instead of 2000) for Israel. They refer to 2005 (instead of 2010) for the Netherlands, to 2008 for Belgium and Iceland, and to 2009 for the Czech Republic and France.

Source: OECD (2012a).

[2] "Addressing the Gender Pay Gap: Government and Social Partner Action", available at http://www.cesifo-group.de/de/ifoHome/facts/DICE/Labour-Market-and-Migration.html.

References

OECD (2008), *OECD Employment Outlook 2008*, OECD Publishing Paris.

OECD (2012a), *OECD Employment Outlook 2012*, OECD Publishing, Paris.

OECD (2012b), *Gender Equality in Education, Employment and Entrepreneurship: Final Report to the MCM 2012*, OECD Publishing, Paris.

European Commission (2006), *The Gender Pay Gap — Origins and Policy Responses* European Commission Directorate-General for Employment, Social Affairs and Equal Opportunities.

Inflows of Asylum Seekers to OECD Countries

Every year hundreds of thousands of people leave their homes to flee international or civil conflicts or persecution of minorities. Most of them come from low or middle income countries, and a large share of the people fleeing persecution seek sanctuary elsewhere in their own country or in nearby countries. Those who end up seeking asylum in OECD countries therefore represent just a small fraction of the people who are displaced against their will.

A key instrument in international refugee policy is the *1951 UN Convention Relating to the Status of Refugees*. Central aspects of the convention still shape refugee policy today. According to the definitions in the convention, a refugee is a person who has fled his or her country or habitual residence because of persecution or a well-founded fear of persecution on account of race, religion, nationality, membership of a particular social group, or political opinion. An asylum-seeker is an individual that claims to be a refugee and applies for sanctuary in a country. Each claim of refugee status must be considered on its individual merits by the signatory country that the asylum seeker has applied to, and the signatory state has to provide access to procedures for determining whether a person claiming asylum qualifies as a refugee according to the Convention's definition.

The United Nations High Commissioner of Refugees (UNHCR) collects comprehensive statistics on refugees and asylum seekers. Table 1 depicts the asylum seeker inflows to 19 leading OECD destination countries in 1989–2010. The largest inflows were received by Germany, the United States, the United Kingdom, France and Canada.

The numbers of asylum seekers reflect conflicts and the humanitarian situation in their countries of origin, but there is no uniform trend in inflows of asylum seekers across different countries. Refugee policies also play a role in determining the different trends between countries. Safe third country provisions represent one such policy.

Safe third country provisions are a form of cost shift and inflow control in OECD countries that follow on from the *1990 Dublin Convention*, and were gradually incorporated into the asylum systems of individual countries. A safe third country is a country that the asylum-seeker has passed through on the way to the receiving country and with which the latter has an agreement. Under an agreement, the receiving country can refuse to examine an asylum application if the country an asylum seeker has passed through is technically responsible for doing so.The purpose of the safe third country policies was to prevent 'asylum shopping', and major recipients of asylum seekers in the EU have generally advocated agreements between the member states to shift some asylum determination responsibilities to other countries.

Safe third country provisions are probably responsible for some of the shift in the distribution of asylum-seekers between countries in the EU, in cases where asylum seekers are more likely to claim asylum from countries that are more easily reachable from outside the EU.

Ilpo Kauppinen

References

OECD, *Trends in International Migration*, SOPEMI 1999, Paris 1999; SOPEMI 2001, Paris 2001; SOPEMI 2002, Paris 2002; SOPEMI 2008, Paris 2008; SOPEMI 2009, Paris 2009, online version (accessed 19 December 2009); SOPEMI 2010, Paris 2010; SOPEMI 2011, Paris 2011.

Table 1

Inflows of asylum seekers, 1989–2010 (in thousands)

	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Austria [a)]	21.9	22.8	27.3	16.2	4.7	5.1	5.9	7.0	6.7	13.8	20.1	18.3	30.1	39.4	32.4	24.6	22.5	13.3	11.9	12.8	15.8	11.0
Belgium	8.2	13.0	15.4	17.5	26.4	14.4	11.4	12.4	11.8	22.0	35.8	42.7	24.5	18.8	16.9	15.4	16.0	11.6	11.1	12.3	17.2	19.9
Czech Republic		1.8	2.0	0.9	2.2	1.2	1.4	2.2	2.1	4.1	7.2	8.8	18.1	8.5	11.4	5.5	4.2	3.0	1.9	1.7	1.4	0.5
Denmark	4.6	5.3	4.6	13.9	16.5	6.7	5.1	5.9	5.1	9.4	7.1	12.2	12.5	6.1	4.6	3.2	2.3	1.9	1.9	2.4	3.8	5.0
France [b)]	61.4	54.8	47.4	28.9	27.6	26.0	20.4	17.4	21.4	22.4	30.9	38.7	54.3	59.0	59.8	58.5	49.7	30.7	29.4	35.4	42.1	47.8
Germany	121.3	193.1	256.1	438.2	322.6	127.2	127.9	116.4	104.4	98.6	95.1	78.6	88.3	71.1	50.6	35.6	28.9	21.0	19.2	22.1	27.6	41.3
Hungary				0.9	0.7	0.2	0.1	0.2	0.2	7.1	11.5	7.8	9.6	6.4	2.4	1.6	1.6	2.1	3.4	3.1	4.7	2.5
Ireland		0.1			0.1	0.4	0.4	1.2	3.9	4.6	7.7	10.9	10.3	11.6	7.9	4.8	4.3	4.3	4.0	3.9	2.7	1.9
Italy [b)]	2.3	4.7	31.7	2.6	1.3	1.8	1.7	0.7	1.9	11.1	33.4	15.6	9.6	16.0	13.5	9.7	9.5	10.3	14.1	30.3	17.6	8.2
Netherlands	13.9	21.2	21.6	20.3	35.4	52.6	29.3	22.2	34.4	45.2	42.7	43.9	32.6	18.7	13.4	9.8	12.3	14.5	7.1	13.4	14.9	13.3
Poland				0.6	0.8	0.6	0.8	3.2	3.5	3.4	3.0	4.6	4.5	5.2	6.9	8.1	6.9	4.4	7.2	7.2	10.6	6.5
Spain [b)]	4.1	8.6	8.1	11.7	12.6	12.0	5.7	4.7	5.0	6.7	8.4	7.9	9.5	6.3	5.9	5.5	5.3	5.3	7.7	4.5	3.0	2.7
Sweden	30.0	29.4	27.4	84.0	37.6	18.6	9.0	5.8	9.7	12.8	11.2	16.3	23.5	33.0	31.3	23.2	17.5	24.3	36.4	24.4	24.2	31.8
United Kingdom [b)]	16.8	38.2	73.4	32.3	28.0	42.2	55.0	37.0	41.5	58.5	71.1	98.9	91.6	103.1	60.1	40.6	30.8	28.3	28.3	31.3	30.7	22.1
Norway	4.4	4.0	4.6	5.2	12.9	3.4	1.5	1.8	2.3	8.4	10.2	10.8	14.8	17.5	16.0	7.9	5.4	5.3	6.5	14.4	17.2	10.1
Switzerland	24.4	35.8	41.6	18.0	24.7	16.1	17.0	18.0	24.0	41.3	46.1	17.6	20.6	26.1	20.8	14.2	10.1	10.5	10.4	16.6	16.0	13.5
Australia	0.5	3.8	16.0	13.4	4.9	6.3	7.6	9.8	9.3	8.2	9.5	13.1	12.4	5.9	4.3	3.2	3.2	3.5	4.0	4.8	6.2	8.3
Canada	19.9	36.7	32.3	37.7	21.1	22.0	26.1	26.1	22.6	23.8	29.4	34.3	44.0	39.5	31.9	25.8	20.8	22.9	28.3	34.8	34.0	23.2
United States [c)]	101.7	73.6	56.3	145.5	200.4	144.6	149.1	107.1	52.2	35.9	32.7	40.9	59.4	58.4	43.3	45.0	39.2	41.1	40.4	39.4	38.1	41.0

Empty cells: Data not available. * Preliminary data a) Excluding de facto refugees from Bosnia Herzegovina. b) Excluding accompanying dependents.
c) Excluding accompanying dependents. Fiscal years (October to September of the year indicated). From 1993 on, figures include applications reopened during year.

Sources: OECD, *Trends in International Migration*, SOPEMI 1999, Paris 1999, p. 263; SOPEMI 2001, Paris 2001, p. 280; SOPEMI 2002, Paris 2002, p. 293; SOPEMI 2003, Paris 2003, p. 306; SOPEMI 2004, Paris 2005, p. 315;SOPEMI 2006, Paris 2006, p. 37; SOPEMI 2007, Paris 2007, p. 321; SOPEMI 2008, Paris 2008 p. 315; SOPEMI 2009, Paris 2009, Online version, accessed 19 December 2009; SOPEMI 2010, Paris 2010, p. 281; SOPEMI 2011, Paris 2011, p. 365.

New at DICE Database

Recent entries to the DICE Database

In the first quarter of 2013 the DICE Database received a number of new entries, consisting partly of updates of existing entries and partly of new topics. Some topics are mentioned below.

- Central banks (organisation, disclosure, accountability, transparency)
- Banking crises responses
- Accounting requirements for SMEs
- Minimum wage
- Labour market efficiency
- Pension reform measures
- Policies to support renewable energies
- Renewable energy targets

Forthcoming Conferences

CESifo Area Conference
on Employment and Social Protection 2013
10–11 May 2013, in Munich

The purpose of the workshop is to bring together CESifo members to present and discuss their ongoing research, and to stimulate interaction and co-operation between them. All CESifo Research Network members are invited to submit their papers, which may deal with any topic within the domains of employment and social protection. Both domains are to be broadly defined, the former including, in particular, issues of the organisation of labour. The latter domain, in turn, includes not only governmental institutions of the welfare state, like social insurance, but also other non-governmental institutions of the welfare society, such as the family, or charities and informal networks, social norms and altruistic behaviour. This conference is open to CESifo network members only.

Scientific organiser: Kai A. Konrad

CESifo Area Conference on Global Economy 2013
17–18 May 2013, in Munich

The conference is intended to allow presentation of current research undertaken by members of the Network's Global Economy area and to stimulate interaction and co-operation between area members. Papers can be on any topic under the Global Economy rubric, covering trade, international finance, migration, global environmental issues, and others. Papers will be discussed in seminar format. Accepted papers will be published as CESifo working papers after revision. This conference is open to CESifo Network members only.

Scientific organisers: Peter Egger and John Whalley

CESifo Venice Summer Institute 2013
22–27 July 2013, in Venice

Five workshops dealing with the following topics:

- The Determinants of Gender Gaps: Institutional Design and Historical Factors
- Emissions Trading Systems as a Climate Policy Instrument: Evaluation & Prospects
- Political Economy and Instruments of Environmental Politics
- The Economics of Language Policy
- The Economics of Infrastructure Provisioning: The (Changing) Role of the State

CESifo Venice Summer Institute is held in co-operation with the Venice International University

New Books on Institutions

Critical Issues in Taxation and Development
Edited by Clemens Fuest and George R. Zodrow,
MIT Press 2013.

Unions, Central Banks and EMU: Labour Market Institutionsvand Monetary Integration
Bob Hancké,
Oxford University Press 2013.

The Welfare State as Crisis Manager
Peter Starke, Alexandra Kaasch and
Franca Van Hooren,
Palgrave Macmillan 2013.

Ifo World Economic Survey

Ifo Institute
Business Surveys Division/WES
Poschingerstr. 5
81679 Munich
Germany

Tel +49-89-9224-1227
Fax +49-89-9224-1911 or
+49-89-9224-1463
E-mail: plenk@ifo.de

WES Membership Form

As a member of the WES Expert Group I shall regularly receive free of charge:

- *CESifo World Economic Survey* quarterly report
- Press releases on the World Economic Survey per e-mail
- the WES one-page questionnaire

Contact person:

Name of institution:

Name of department:

Address:

Country:

Telephone:

Fax:

E-mail:

Thank you for your co-operation!

DICE

Database for Institutional Comparisons in Europe

www.cesifo-group.org/DICE

The DICE database was created to stimulate the political and academic discussion on institutional and economic policy reforms. For this purpose, DICE provides country-comparative information on institutions, regulations and the conduct of economic policy.

To date, the following main topics are covered: Business and Financial Markets, Education and Innovation, Energy and Natural Environment, Infrastructure, Labour Market and Migration, Public Sector, Social Policy, Values and Other Topics.

The information of the database comes mainly in the form of tables – with countries as the first column – but DICE contains also several graphs and short reports. In most tables, all 27 EU and some important non-EU countries are covered.

DICE consists primarily of information which is – in principle – also available elsewhere but often not easily attainable. We provide a very convenient access for the user, the presentation is systematic and the main focus is truly on institutions, regulations and economic policy conduct. Some tables are based on empirical institutional research by Ifo and CESifo colleagues as well as the DICE staff.

DICE is a free-access database.

Recommendations are always welcome.
Please address them to

poutvaara@ifo.de
or
DICE@ifo.de