



norden

NORDIC ECONOMIC POLICY REVIEW

CHALLENGES IN HEALTH CARE FINANCING AND PROVISION

Introduction

Tor Iversen and Sverre A.C. Kittelsen

Ageing populations: More care or just later care?

Terkel Christiansen, Jørgen Lauridsen and Mickael Bech

Lifestyle, health and costs – what does available evidence suggest?

Kristian Bolin

The economics of long-term care: A survey

Helmuth Cremer, Pierre Pestieau and Gregory Ponthiere

The role of primary health care in controlling the cost of specialist health care

Stephen Beales and Peter C. Smith

Payments in support of effective primary care for chronic conditions

Randall P. Ellis and Arlene S. Ash

An economic assessment of price rationing versus non-price rationing of health care

Luigi Siciliani

Should pharmaceutical costs be curbed?

Kurt R. Brekke, Dag Morten Dalen and Steinar Strøm

Productivity differences in Nordic hospitals:

Can we learn from Finland?

Clas Rehnberg and Unto Häkkinen



The *Nordic Economic Policy Review* is published by the Nordic Council of Ministers and addresses policy issues in a way that is useful for in-formed non-specialists as well as for professional economists. All articles are commissioned from leading professional economists and are subject to peer review prior to publication.

The review appears twice a year. It is published electronically on the website of the Nordic Council of Ministers: www.norden.org/en. On that website, you can also order paper copies of the Review (enter the name of the Review in the search field, and you will find all the information you need).

Nordic Economic Policy Review

Challenges in health care financing and provision

Tor Iversen and Sverre Kittelsen (Editors)

TemaNord 2013:514

Nordic Economic Policy Review
Challenges in health care financing and provision
Tor Iversen and Sverre Kittelsen (Editors)

ISBN 978-92-893-2496-0
<http://dx.doi.org/10.6027/TN2013-514>
TemaNord 2013:514

© Nordic Council of Ministers 2013

Cover photo: Pub. Unit/NCM

Print: Rosendahls-Schultz Grafisk
Copies: 40

Printed in Denmark



This publication has been published with financial support by the Nordic Council of Ministers. However, the contents of this publication do not necessarily reflect the views, policies or recommendations of the Nordic Council of Ministers.

www.norden.org/en/publications

Nordic co-operation

Nordic co-operation is one of the world's most extensive forms of regional collaboration, involving Denmark, Finland, Iceland, Norway, Sweden, and the Faroe Islands, Greenland, and Åland.

Nordic co-operation has firm traditions in politics, the economy, and culture. It plays an important role in European and international collaboration, and aims at creating a strong Nordic community in a strong Europe.

Nordic co-operation seeks to safeguard Nordic and regional interests and principles in the global community. Common Nordic values help the region solidify its position as one of the world's most innovative and competitive.

Nordic Council of Ministers

Ved Stranden 18
DK-1061 Copenhagen K
Phone (+45) 3396 0200

www.norden.org

Content

Challenges in health care financing and provision <i>Tor Iversen and Sverre A.C. Kittelsen</i>	7
Ageing populations: More care or just later care? <i>Terkel Christiansen, Jørgen Lauridsen and Mickael Bech</i>	23
Comment by <i>Anna Lilja Gunnarsdottir</i>	55
Lifestyle, health and costs – what does available evidence suggest? <i>Kristian Bolin</i>	59
Comment by <i>Tinna Laufey Ásgeirsdóttir</i>	99
The economics of long-term care: A survey <i>Helmuth Cremer, Pierre Pestieau and Gregory Ponthiere</i>	107
Comment by <i>Þórólfur Matthiasson</i>	149
The role of primary health care in controlling the cost of specialist health care <i>Stephen Beales and Peter C. Smith</i>	153
Comment by <i>Helgi Tómasson</i>	187
Payments in support of effective primary care for chronic conditions <i>Randall P. Ellis and Arlene S. Ash</i>	191
Comment by <i>Jørgen T. Lauridsen</i>	211

An economic assessment of price rationing versus non-price rationing of health care

<i>Luigi Siciliani</i>	213
Comment by <i>Mickael Bech</i>	243

Should pharmaceutical costs be curbed?

<i>Kurt R. Brekke, Dag Morten Dalen and Steinar Strøm</i>	247
Comment by <i>Helgi Tómasson</i>	275

Productivity differences in Nordic hospitals:

Can we learn from Finland?

<i>Clas Rehnberg and Unto Häkkinen</i>	277
Comment by <i>Thorvaldur Gylfason</i>	317

Challenges in health care financing and provision

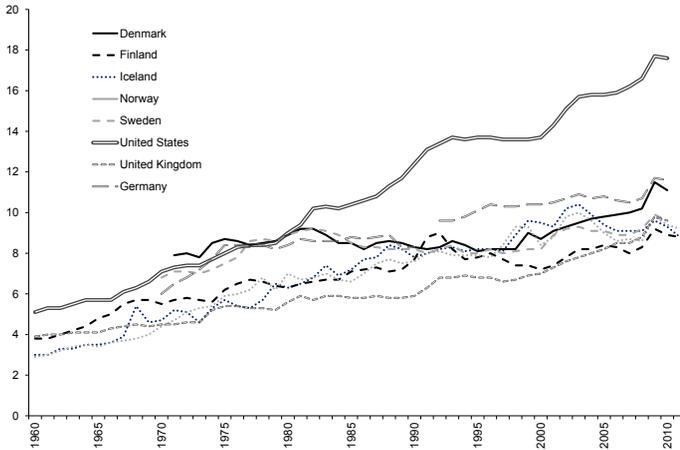
Tor Iversen* and Sverre A.C. Kittelsen**

Good health is highly valued and a prerequisite for taking full benefit of a rising level of income. Hence, the willingness to pay for improved health is likely to rise sharply with income, and so is the willingness to pay for health care, since health care is a vital input in the production of health. Hall and Jones (2007) claim that as people get richer and consumption rises, the marginal utility of consumption falls rapidly. Furthermore, the marginal utility of life extension does not decline and spending on health to extend life allows individuals to purchase additional periods of utility. As a result, the optimal composition of total spending shifts toward health, and the health share grows along with income. In projections based on their quantitative model, they find that the optimal health share of spending seems likely to exceed 30 percent in the US by the middle of the century. This is consistent with the development of health care spending as a percentage of GDP illustrated by Figure 1. All Western countries have had an increase in the share of GDP used on health care since the 1960's. While the US has continued on a rising trend, the share seems to have more or less stabilized in the Nordic and other European countries since the 1990's. Note that the figure does not show the actual level of health care received by the citizens, since the countries also vary in income, cost level and the extent of informal care.

* Department of Health Management and Health Economics, University of Oslo, tor.iversen@medisin.uio.no.

** Ragnar Frisch Centre for Economic Research, Sverre.kittelsen@frisch.uio.no.

Figure 1. Health expenditure as percentage of GDP, Nordic and selected non-Nordic countries



Source OECD (2012).

There are a number of trends that influence health care expenditure besides income. On the demand side, there are the demographic changes that are summarized as an ageing population due to higher life expectancy and changing fertility, changes in lifestyle that may lead to obesity and other risk factors, and environmental changes that may influence health. On the supply side, new and often expensive medical technology gives more and better treatment, and increasing female labour force participation has reduced informal care given in the family. Even though the health care needs are changing and new technologies improve human happiness and welfare, the expansion of health care still needs to be financed. The characteristics of the markets for health insurance and health care make the expansion of health care challenging. These characteristics determine the trade-offs between various objectives and overall goals in the health sector and constitute health economics as a separate field of applied economics. The authors of this issue of the *Nordic Economic Policy Review* have been invited to deal with some of the main challenges in the financing and organisation of health care in a Nordic setting.

Risk aversion and uncertainty about future health imply demand for health insurance to cover future costs of health care. The purpose of health insurance is to relieve the citizens from bearing the financial risk of major health expenses. Hence, health insurance implies that there is a

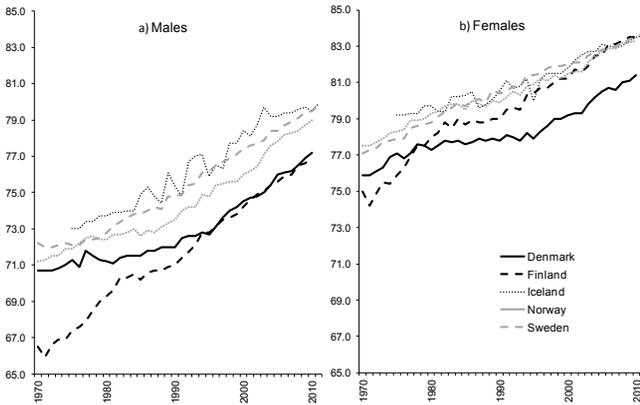
third party that pays for health services. Such third-party financing characterizes all insurance policies, whether public or private, and represents no efficiency problem in itself. The potential efficiency problem arises when information about disease prevention, disease risk, cost and quality of care is unevenly distributed between the three parties; i.e. the patient, the insurance company and the health service provider. For instance, health insurance implies moral hazard and reduced incentives for disease prevention, as described by Kristian Bolin. Variation in disease risk and the degree of risk aversion in the population combined with insurance companies knowing less about the individuals' disease risks than individuals themselves, result in market imperfections due to adverse selection and not everyone may get the insurance coverage they want. Hence, the unequal distribution of information can entail an argument for mandatory insurance. Mandatory insurance with income-dependent premiums can also be justified by the median voter's interests. Uncertainty about future risk groups as well as health-related altruism contributes to a more robust public financing. Public funding will be harder to maintain the greater variability there is in disease risk, the greater the proportion of the population that is at high risk of disease, and the more costly the diagnosis and treatment of disease. The role of the public sector in health care insurance and provision is significant in all developed countries and more prominent in the Nordic countries than in most other countries.

Small patient co-payments imply that patients will demand health care even if their valuation of the marginal health improvements from care is less than the marginal cost of providing health care. Hence, the price mechanism will not fulfil its role of allocating resources to and within the health sector. There is a need for other types of rationing in addition to the limited rationing by means of the price mechanism. Waiting times, implicit prioritizing by service providers and governments' explicit prioritizing are used as rationing mechanisms in the Nordic countries. Luigi Siciliani elaborates on the optimal balance between these instruments for rationing health care in his article.

The relationship between insurer and health service provider is characterized by the insurer having less access to information about the service provider's operations than the service provider itself. This applies to information about the service provider's efforts to reduce costs, information about the patient composition, information about the possible

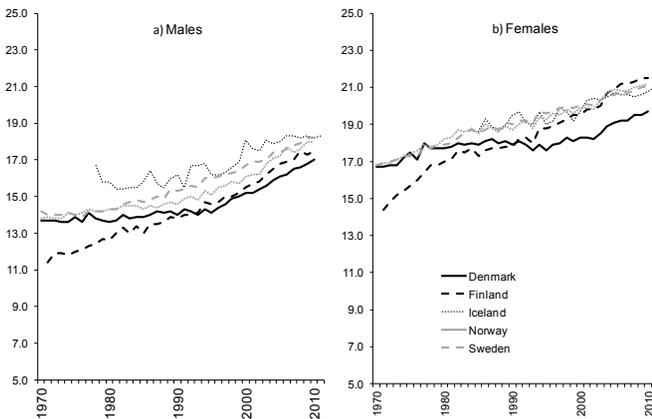
patient selection and information about the quality of care provided. The problems of asymmetric information for the insurer are raised from several perspectives by Rehnberg and Häkkinen (hospitals), Brekke, Dalen and Strøm (pharmaceuticals), Ellis and Ash (risk adjusted capitation payment), Beales and Smith (primary care and specialist care) and Cremer, Pestieau and Ponthiere (long-term care).

Figure 2. Life expectancy at birth, Nordic countries



Source: OECD (2012).

Figure 3. Life expectancy at 65, Nordic countries



Source: OECD (2012).

A growing proportion of old people in the population is the main driver of health care demand besides income. This development has its background both in previous and present birth rates and in the development of longevity. During recent years, there has been a remarkable increase in the longevity of old people in the developed part of the world. Figure 2 illustrates how life expectancy has increased by 5-10 years in all Nordic countries over the past four decades. Eggleston and Fuchs (2012) show that at the beginning of the twentieth century, in the United States and other countries at comparable stages of development, most of the additional years of life were realized in youth and working ages; and less than 20 per cent were realized after the age of 65. Now, they find that more than 75 per cent of the gains in life expectancy are realized after the age of 65 – and that share is approaching 100 per cent asymptotically. They assert that the new demographic transition is a longevity transition and ask how individuals and societies will respond to mortality decline when almost all of the decline will occur late in life. In the Nordic countries, the increase in life expectancy at the age of 65 shown in Figure 3 is particularly noticeable after 1990. Christiansen, Lauridsen and Bech elaborate on an important part of this question in this issue by asking "Ageing populations: More care or just later care?". Cremer, Pestieau and Ponthiere take the issue further with their survey of the economics of long-term care for the elderly.

The development of unhealthy lifestyles in rich societies has recently attracted much attention. In particular, the growing occurrence of obesity is a concern both because of its negative health effects and its potential effects on the demand for health care. Lifestyle and the related health and cost constitute the topic of Kristin Bolin's paper. Tinna Laufey Ásgeirsdóttir elaborates on the issues related to the deviance from rational behaviour in her comment to Bolin's paper.

The development in medical technology has contributed to increased longevity and improved quality of life (see, for instance, Cutler, 2004 for examples). The development in medical technology has also contributed to increased costs of health care, since patients who were previously offered no care or inferior care are now offered more effective care that helps them survive. In public insurance systems, it becomes a social concern what level of costs that is acceptable for obtaining a marginal gain in longevity and quality of life, and what instruments that are available for

implementing the socially optimal amount and composition of health care. These issues are raised at a fundamental level by Siciliani, specifically for pharmaceuticals by Brekke, Dalen and Strøm and for hospitals by Rehnberg and Häkkinen. Rehnberg and Häkkinen are in particular interested in what role comparative studies can have as an instrument for raising hospital productivity.

Taken together, the ageing population, development in medical technology and unhealthy life styles give rise to an increasing *occurrence of chronic diseases*. While patients previously died of their diseases, they now live with their diseases for more years. Patients with chronic diseases typically demand health care from many types of providers. Hence, their access to health care and the coordination of their many different needs for health care is a crucial issue. Beales and Smith review and discuss the literature on whether or not primary care can take care of patients at a lower cost than specialist care. Ellis and Ash describe and discuss how payment systems for patients with chronic diseases can be constructed to avoid poor access for patients who need many services.

1. Ageing populations: More care or just later care?

The increasing share of elderly in European countries stems partly from increased life expectancy and partly from decreasing fertility among the younger generations. The effect of ageing on health care expenditure depends crucially on the development of the health of the elderly, as well as on other factors such as medical technology and the institutions of the health care systems in each country. It is generally observed that health care utilization increases with age, and a large proportion comes at the end of life. In a pessimistic and costly scenario, increasing life expectancy will not increase the healthy lifespan but will just expand the period of ill-health at the end of life. At the other extreme, health technology improvements may imply shorter periods of ill health at the end of life and therefore lower expenditure for each person. In such a scenario, health care expenditures need not increase even if the share of elderly does.

Terkel Christiansen, Jørgen Lauridsen and Mickael Bech survey the studies that have been made on the expenditure effect of ageing. In studies based on data on individuals from a single country, it is possible to

distinguish between health care costs that increase with age in general and those that are due to proximity to death. The results in these studies are mixed, but while proximity to death generally incurs large costs, many studies only find modest increases in age-specific costs. Some even find declining age-specific costs for the very old above e.g. 80 or 85.

Studies based on individuals in one country have difficulties in capturing the effect of life expectancy, or of different institutional arrangements, at least in the short term. Instead, one can use comparisons of countries over several years. In these studies, income is always the main determinant of health care expenditure, although health care expenditure is publicly regulated in most countries. While different studies show different results due to differences in methods and data, the general impression is that ageing as such can be expected to cause only a modest increase in health care expenditure per capita in the future. This conclusion is supported by the authors' own empirical study, based on 15 EU countries.

2. Lifestyle, health and costs – What does available evidence suggest?

In this article, Kristian Bolin provides a summary of what is known regarding (1) health risks, and healthcare and productivity costs, and (2) the effectiveness and cost-effectiveness of primary and secondary prevention programmes associated with smoking, alcohol abuse, nutritional choices, physical activity, and obesity. Health risks associated with smoking are well-established and quantitatively large compared with other health risks included in the study. Consequently, healthcare costs and costs related to productivity that can be attributed to smoking are fairly well-known, and both primary and secondary effective and cost-effective preventive interventions against smoking are available. The health risks associated with alcohol consumption are also considerable – although not as high as those for daily smokers – for those who consume excessive amounts. The risks decrease with consumption and some studies have even found beneficial health effects associated with moderate consumption. Alcohol-attributable healthcare and productivity costs are also relatively well-known. Some evidence suggests that primary and secondary alcohol pre-

vention may be both effective and cost-effective. The evidence, however, is less reliable than the corresponding evidence for smoking.

The health risk, and healthcare and productivity costs, associated with inadequate physical activity and obesity are fairly well-known. However, healthcare and productivity costs associated with specific dietary patterns, and specific foods, are disputable, due to the unclear relationships between diet and future health outcomes. The cost-effectiveness of interventions designed for changing health behaviours is largely unknown. Both primary and secondary anti-obesity programmes that target child and adolescent behaviour are potentially effective.

3. The economics of long term care: A survey

This article by Helmut Cremer, Pierre Pestieau and Gregory Ponthiere presents an overview of the economic literature on long-term care. It first presents some evidence on the extent of the problem of disability in very old age. With an ever-increasing longevity, the needs are expected to increase in most countries and this exerts pressure on the three institutions providing LTC: the family, which is by far the main provider, the market and the state. Each of these institutions raises specific problems. The role of the family is under pressure because of the increasing rate of labour market participation among women aged fifty and plus, because of the increasing mobility of children and because of changes in family values. The private market for LTC insurance remains very thin almost everywhere. This is due to a variety of reasons that are both economic and social. Finally, in most countries, the state remains reluctant to offer a universal social insurance programme for long-term care that would advantageously replace the current, often quite cumbersome means-tested systems.

The main lesson emerging from the overview is that there is a great deal of interaction between these three institutions, the most noticeable being the partial crowding out exerted by public programmes on family solidarity and on the private insurance market. Another lesson is that, even if it were possible, returning to the old scheme of family solidarity is not necessarily desirable as it can hide situations of forced solidarity in which case human and social costs can be huge.

In the discussion of policy alternatives, the authors distinguish between direct and indirect involvement of the state. Directly, the state faces a tough choice between a fully-fledged universal coverage social insurance and a means-test programme restricted to the poorest. The first type of system is much more expensive. However, the second can only be effective if the means testing is rigorously enforced (which is currently not the case in many places). The authors argue that public action can be useful to foster the LTC insurance market and to keep family and community solidarity as active and effective as possible. Mere tax incentives would be insufficient. The government should provide education and information on the risks of dependence and the type of services that each type of dependence requires. Many people seem to be unprepared for the risks of dependence, in the same way in which they were unprepared for the risk of retirement half a century ago.

4. The role of primary health care in controlling the cost of specialist health care

The motivation for the article by Stephen Beales and Peter C. Smith is the concern in developed health systems that increases in health expenditure have reached unsustainable levels, leading to an urgent search for expenditure control mechanisms. One particular concern is the use of hospital inpatient services that is supposed to be ‘avoidable’ in most OECD countries. According to the authors, the belief is that – with timely, high quality intervention in primary care – unnecessary specialist health care utilization could be markedly reduced, with associated cost savings and improved quality of life for patients. The authors review the empirical evidence for three broad forms of primary care intervention: reducing or delaying the onset of disease; reducing the use of specialist care once a clinical condition has been identified; and reducing the intensity of use of specialist care once a need for such care has arisen. They find little persuasive evidence on the macro benefits of primary care spending in terms of reduced overall spending, and – with a few exceptions – the micro evidence is small scale and inconclusive, although there are indications of promising policy options for future experimentation. Then, they examine the role of incentives in promoting the cost containment role of primary

care. In general, the empirical results of experiments are found to be disappointing. The paper concludes with a discussion of why this might be the case and the associated policy implications. They state that the experience in all health system reforms is that a reform stands little chance of success without clinical leadership and engagement, including at the most senior level. They also state that disappointing results from some pilot schemes may be due to their small scale, or the short time for which they are implemented. Finally, the authors note that the distinction between primary and secondary care may become increasingly blurred in future years. As the number of older people with complex chronic medical needs increases, so does the demand for integration of care, and personalized medical treatment will grow. Whether there will be a provider response to such demand is likely to depend on the reform of provider payment mechanisms, particularly for secondary care. At present, these usually reward discrete episodes of care. In the future, payment mechanisms are increasingly likely to reward 'bundles' of care, or indeed a whole year of care, for people with complex needs. Experience in the US with the new "Accountable Care Organizations", responsible for the costs and quality of health care for a defined population (with a minimum size of 5 000 people), will be of great interest in this respect.

5. Payments in support of effective primary care for chronic conditions

How to appropriately reward bundles of care is the topic of the paper by Randall P. Ellis and Arlene S. Ash. When bundled payments are large, weak risk adjustment creates a strong incentive for practices to avoid individual patients expected to cost more than the bundled payment. Hence, there is a danger that the chronic patients in most need of care are the group that receives the poorest quality of care. Perverse economic incentives might make the group that is the focus of a health care reform end up as losers. Ellis and Ash describe and discuss risk adjustment based on their work in developing efficient risk adjustment systems in the US. The experiences have great relevance for the Nordic countries in their ambitions to coordinate health care between primary care and specialist care. Ellis and Ash assert that risk adjustment models can be calibrated

and used to establish appropriate payments and incentives for delivering superior primary care, particularly to people with chronic conditions requiring careful management, through health-based capitation payment and performance assessment in a patient-centred medical home (PCMH). The implementation of risk-adjusted primary care payment for a PCMH will be easier in Scandinavian countries where payments are made by a single payer; however, the decentralized administrators responsible for paying for primary care may face many of the same challenges that appear in the US. They address practical considerations and administrative structures that could support a risk-adjusted payment reform for the PCMH. Feasibility is supported by the experience of one health plan in the US that conducted a “virtual all-payer” PCMH pilot. Their approach could serve as an inspiration also for policy-makers and health care administrators in the Nordic countries.

6. An economic assessment of price rationing versus non-price rationing of health care

Although health insurance, public or private, aims at reducing the costs to the patient if and when the need arises, it is generally assumed that if health care was freely available, utilization would be higher than optimal and the expenses would be excessive. Thus, health care has to be rationed in one way or another. This article by Luigi Siciliani reviews the relative merits of three different forms of rationing: i) price rationing, which takes the form of a co-payment or a coinsurance rate, and two forms of non-price rationing, ii) rationing by waiting, when a patient is placed on a waiting list before receiving treatment, and iii) explicit rationing, when the patient is explicitly refused treatment. Both waiting times and co-payments can help contain excess demand, though the demand is generally inelastic with respect to waiting times and co-payments (elasticities of -0.1 or -0.2).

With price rationing, the patient or a physician, acting as the patient's representative, weighs the benefits of the health care against the price, a mechanism which in most markets leads to a socially optimal use of the service. This does, however, defeat the very objective of health insurance, which is to reduce out of pocket expenses when the need arises. In addi-

tion, it is inequitable if the use of health care depends more on patients' income than on their real need. Information problems imply that it is not always easy for the patient to know the true benefit.

Rationing by waiting time in public health care works by either shifting some patients to the private sector, by deterring doctors from referring patients to treatment, or by making patients give up waiting. While the first mechanism may be good for redistribution since the rich pay for their own treatment, the informational problems mean that it may not be the patients with the lowest needs that are rationed. Most importantly, while price rationing incurs a cost to the patient but gives income to the provider or insurer, waiting time incurs a real cost for the patients without giving a benefit to anyone.

Explicit rationing can potentially generate higher patient welfare than co-payments or waiting times, since there are no price or waiting costs for the patients. If doctors are able to assess the health needs of the patients, then treatment is given to those with the highest benefit. Explicit rationing can take the form of a list of treatments that are not covered by the public health insurance, but more often one needs to set a threshold so that some patients get treatment and some do not. This may be costly for the doctor, who has to act as a gatekeeper, but the costs to the doctor can be reduced by more precise clinical guidelines. The author recommends an increased development of such guidelines to facilitate more use of explicit rationing.

7. Should pharmaceutical costs be curbed?

Both pharmaceutical innovations and the ageing of the population explain the increasing importance of pharmaceuticals in health care, which is here discussed by Kurt Brekke, Dag Morten Dalen and Steinar Strøm. Pharmaceuticals account for almost a fifth of total health spending in OECD-countries. However, in the Nordic countries, the expenditure has stabilized over the past few years, especially in Norway where the expenditure has not increased since 2004. There are considerable differences in both expenditure and price level between countries.

Due to the importance of patent protection and insurance coverage, pharmaceutical markets are subjected to economic regulation – both on

the supply side and the demand side. This article explains the special features of pharmaceutical markets and the Nordic markets in particular, before explaining the main regulatory policy measures taken by governments in these countries. To encourage the development of new drugs, patents protect the innovating company from direct competition. Policy instruments are important in avoiding excessive use or pricing of patented drugs, but also in encouraging effective competition after the patent expires. A large proportion of drug expenditure is paid by the public sector either through hospital budgets or health insurance schemes.

Demand can be regulated by co-payments or co-insurance from patients. Reference pricing is a co-payment scheme that has become increasingly popular in recent years. The regulator sets a reference price, which is the maximum reimbursable price for all drugs in the reference group. Direct regulation, akin to explicit rationing, requires that a drug meets a minimum cost-efficiency ratio if the regulator is to place the drug on the reimbursement list.

On the supply side, prices can be regulated both by price caps on wholesale or end user prices, and on the mark-ups at the retail level. Price caps need direct or indirect information on the cost level of the suppliers, and international price comparisons are often used effectively. Generic substitution allows or requires pharmacies to substitute a prescribed brand-name drug with a cheaper generic version with the same substance or the same therapeutic effect.

In empirical studies reviewed in the article as well as in studies conducted by the authors themselves, there is evidence that economic regulation does work. The authors point out that using cost-effective drugs benefits patients and increases social welfare. The large increases in drug expenditure in the 1990's were strongly influenced by the introduction of new and innovative drugs with new benefits, and the flattening expenditure curve in recent years may be evidence of more mature markets and effective regulatory policies. In addition to price-lowering policies, the authors emphasize the importance of cost-efficiency or cost-benefit analysis when drugs are approved for reimbursement.

8. Productivity differences in Nordic hospitals: Can we learn from Finland?

Comparative studies of health system performance are a source for identifying and explaining differences in costs, outcome and efficiency. Acute short-term hospitals are the major resource users in the health care sector and have a significant role for advanced treatment. In this paper, Clas Rehnberg and Unto Häkkinen present and discuss the findings from the Nordic collaboration on productivity differences across acute hospitals. As the four countries share many administrative tools and use common standards for data collection, unique cross-country comparisons are possible. The results suggest that there is a markedly higher average hospital productivity in Finland compared with Denmark, Norway and Sweden. Further analysis shows country-specific effects not to be correlated with the explanatory variables tested. This means that these country effects must be linked to the structure of financing, regulatory framework, organisational arrangements etc. in each country.

The explanations of findings are discussed along with different theories and possible reasons for the observed differences. Although no clear explanations are argued for, a number of hypotheses for further research are identified. The markedly higher productivity levels among the Finnish hospitals do not seem to be explained by differences in the use of market mechanisms and reimbursement systems. The Finnish system has not implemented performance-based payments or internal market mechanisms. The method and arrangements for the allocation of resources in Finland between different health services, as well as the trade-off against other public sector tasks at the municipality level, are proposed as major differences in relation to the neighbouring countries. The combined role as purchaser and provider at the municipality level is also proposed as important for the resource allocation within the health sector. The paper argues for a closer analysis of the impact of fund-holding, contractual relations and incentives between levels of governments as well as including quality indicators in the efficiency measure.

9. Lessons to be learned

Much of the growth in health expenditure stems from increased income, from the increase in life expectancy and the share of the elderly, and from the availability of new treatments. These factors imply that a continued growth in health expenditure may increase social welfare. Some measures can be explored to decrease the burden of financing, and to improve welfare for a given level of expenditure.

In the long run, there is the challenge of designing migration and fertility policies that may stabilise the demographic composition of the population. In the medium run, the demand for health care can be crucially influenced by the encouragement of a healthy life style. As discussed by Bolin and by Ásgeirsdóttir, it is not clear how effective policies should be designed. These questions clearly underline the need for field experiments and further research more generally.

There seems to be more scope for policy on the supply side. Brekke, Dalen and Strøm show how economic regulation of the pharmaceutical markets has an effect by lowering prices. Rehnberg and Häkkinen demonstrate that there are considerable differences in the productivity of Nordic hospitals, pointing the way to potential cost savings. The reasons for these differences are not fully understood, and should once more be the subject of further research. One possible explanation is that the more productive Finnish hospitals are owned by municipalities that also have the responsibility for primary care and for other public expenditure areas such as education.

This is in accordance with the discussions of Beales and Smith and of Ellis and Ash who point to the importance of increased weight on integrated care, particularly for chronic patients. These same authors also emphasise the importance of designing reimbursement schemes that reflect the risk profile of the population.

Perhaps the clearest policy implication of the articles in this issue of *NEPR* is the need for more explicit priorities as to which patients should be treated. Siciliani argues strongly that explicit rationing has less efficiency loss than rationing by waiting times and is more equitable than price rationing. As Brekke, Dalen and Strøm also point out, treatments and drugs that are cost-efficient should be financed, but those that do not have a minimum effect for the cost incurred should not be financed. To

be effective and equitable, these decisions must be based on health economic evaluations, clinical guidelines and priorities decided at a central level.

There have been constraints of width and depth in the selection of themes for this issue of NEPR. Closely related issues such as pension policy are not covered, since this is beyond the field of health economics. Not treated in this issue of the *Review* is the major question of the organisation of health insurance as such: Is there a place for a larger involvement of private health insurance, such as in the mandatory insurance scheme of the Netherlands? We have taken the Nordic model as given, and invited authors to discuss challenges within that context.

Brave reform proposals in the health care sector are often initiated with scant knowledge of their effects. This is in remarkable contrast to the documentation that new pharmaceuticals and medical procedures are required to deliver. There is a need for more experiments also in the organisation and financing of the health sector. The Nordic countries are in a favourable position for doing this kind of research with standardized register data covering the entire population. Comparative studies in the Nordic setting have a potential for contributing to an improved knowledge basis of health policy.

References

- Cutler, D.M. (2004), *Your Money or Your Life: Strong Medicine for America's Health Care System*, Oxford University Press, New York.
- Eggleston, K.N. and Fuchs, V.R. (2012), The new demographic transition: Most gains in life expectancy now realized late in life, *Journal of Economic Perspectives* 26, 137-156.
- Hall, R.E. and Jones, C.E. (2007), The value of life and the rise in health spending, *Quarterly Journal of Economics* 122, 39-72.
- OECD (2012), OECD health data 2012 – Frequently requested data, <http://www.oecd.org/health/healthpoliciesanddata/oecdhealthdata2012-frequentlyrequesteddata.htm> (October 21, 2012).

Ageing populations: More care or just later care?*

Terkel Christiansen **, Jørgen Lauridsen ***, and Mickael Bech ****

Summary

An ageing society is characterised by an increasing median age of the population. The purpose of this chapter is to document the existing knowledge about the association between population ageing and health care expenditure, and to supplement this overview by a summary of our original research. While different studies show different results due to differences in methods and data, the general impression is that ageing as such can be expected to only cause a modest increase in health care expenditure per capita in the future. This conclusion is supported by our own empirical study, based on 15 EU countries.

Keywords: health care expenditure, ageing.

JEL classification numbers: H51, J11, J14.

* The project has received financing from the European Commission under the 6th Research Framework Program (Contract No. SP21-CT-2003-502641).

** Centre of Health Economics Research (COHERE), Department of Business and Economics, University of Southern Denmark, tch@sam.sdu.dk.

*** Centre of Health Economics Research (COHERE), Department of Business and Economics, University of Southern Denmark, jtl@sam.sdu.dk.

**** Centre of Health Economics Research (COHERE), Department of Business and Economics, University of Southern Denmark, mbe@sam.sdu.dk.

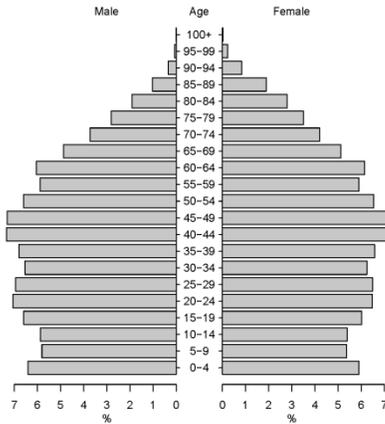
Population ageing is a process that takes place when the median age of a population is increasing, due to increasing life expectancy, decreasing fertility or both. Ageing is a common feature of all European countries and derives from a decrease in both mortality and fertility as well as an increase in life expectancy. When individuals desire to have a long life and a high quality of life, this presents a challenge to society due to the current setup of the public benefits and the retirement age. It has thus been estimated that the ratio of the population above 65 to the population in labour active groups will increase from 24 to 49 per cent within the old EU-15 countries between 2000 and 2040 (Morrow and Roeger, 1999). Such demographic changes can be foreseen to have far-reaching consequences for total economic output and hence, total welfare, public budgets, income distribution, the labour market and financial markets, and maybe even social coherence.

This type of concern is not new, however. Already in 1934, Gunnar and Alva Myrdal (1934) addressed the problem in their book *Kris i Befolkningsfrågan (Crisis in the Population Question)* due to the decreasing fertility in their country (Sweden) during the economic depression at that time. In 1940, Gunnar Myrdal (1940) expressed concern about the economic consequences of decreasing fertility. Later, the fertility trend was circumvented by the post-war baby boom, and attention was drawn away from the issue until a new decline in fertility occurred in later decades along with a steady decrease in mortality and an increase in life expectancy (Hagemann and Nicoletti, 1989).

Figures 1 and 2 show age pyramids for Northern Europe in 2011 and 2050. An age pyramid is used to show the distribution of age and gender in a population at a given point in time. Obviously, the traditional pyramid-like shape of such figures has changed, and a substantially larger share of the population is predicted to be above 65 in 2050 as compared to 2011 (UN, 2011).

For the health care sector in particular, concerns have also been voiced about increasing health care expenditure (HCE) due to an ageing population, and a large number of empirical studies have addressed the problem. However, no clear-cut conclusions as to the effect of population ageing on HCE per capita have been made due to differences in methods and data availability.

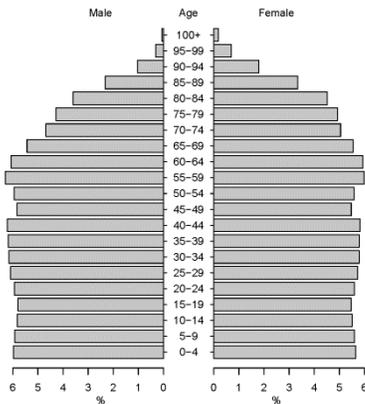
Figure 1. Age pyramid 2011, Northern Europe



Source: UN (2011).

Note: Includes Channel Islands, Denmark, Estonia, Finland, Iceland, Ireland, Latvia, Lithuania, Norway, Sweden, UK.

Figure 2. Age pyramid 2050, Northern Europe



Source: UN (2011).

Note: see note to figure 1.

It is a popular belief that because individual average health care costs generally increase by age, as can be documented by cross-sectional data, the total health care costs per capita will inevitably increase due to an

ageing society. Previous studies have not unanimously supported this conclusion, however, due to the lack of accounting for the observed high health care costs in a short time period prior to death. As the population ages, these costs are postponed by the postponement of deaths. In essence, predictions from cross-sectional data on health care utilisation involve a fallacy because population health dynamics must be taken into account.

It is quite another issue whether or not an increasing share of the population being elderly will incur an extra health care “burden” upon the younger generation through, e.g., increasing taxation or payment of insurance premiums to health care.

In the present article, population ageing refers to an increasing share of the population being aged above 65. The literature we review sometimes uses the terms expenditure and costs as synonyms. In general, we use the term *costs* when we refer to the use of resources for individuals and *health care expenditure* (HCE) when we refer to society’s budget allocation to the health care sector.

The chapter will focus on Western countries in general and selected EU countries in particular. The purpose of this chapter is two-fold: 1) to document the existing knowledge about the association between population ageing and HCE on the basis of an overview of the literature, and 2) to supplement this overview by our original research.

1. Ageing scenarios and health care costs

Studies on ageing have applied various scenarios for longevity and health status with consequences for HCE. In a simplified *status quo* scenario, the age-specific costs only depend on the state of medical technology (Breyer and Felder, 2006). Thus, when technology is assumed to be unchanged, future HCE can be estimated by using present age-specific cost and population projections for a prognosis of future HCE. In the *compression of morbidity* scenario, it is assumed that lifetime in good health increases, whereas life expectancy is constant, which implies a shorter time living with morbidity (Fries, 1980). In contrast, in the *expansion of morbidity* scenario, it is assumed that the age at the onset of morbidity is constant but life expectancy increases, which implies a longer time living with

morbidity (Kramer, 1980; Gruenberg, 1977). This is a pessimistic scenario, as it implies that while it is possible to cure certain diseases and prevent people from dying, new diseases will occur during the prolonged lifetime, which will incur extra costs. Moreover, the average health status of the surviving population will decrease (Breyer and Felder, 2006).

Various combinations of these scenarios have been formulated: one is a *healthy ageing* scenario with increased lifetime in good health and increased life expectancy (Manton, 1982). A general theory of population ageing was formulated by Michel and Robine (2004). A modification of the compression of the morbidity scenario is that health care costs may even decline above a certain age (Lubitz et al., 1995). Possible reasons might be a reluctance to treat very old terminally ill patients, or the use of long-term care rather than hospitals.

It has been observed in a range of studies (see Section 2.1 below) that individual health care costs are particularly high during a short period before death, which means that proximity to death matters. When the mortality risk shifts upward in age groups, costs at the end of life will occur later but they will not affect individual lifetime health care utilisation. Still, health care costs may increase gradually by age (over a certain age) for an individual. Hence, if the healthy ageing scenario is accepted, the health care costs will increase more slowly over time than what might be expected from an observation of present-time individual cost profiles by age.

2. Review of the literature

A large number of empirical studies from the last two decades or more have examined the determinants of HCE, including ageing. When reviewing the literature, a distinction can be made between studies that are based on individuals (micro-level studies) and analyses that have countries as the unit of observation (macro-level studies). Since micro-level studies are usually based on single countries or a region within a country, only variables related to individuals are included (demand-side variables). In contrast, macro-level studies allow for the inclusion of variables that characterise each country's health care system (supply-side variables). Both types of studies are briefly reviewed below. It appears from the

review that the studies vary significantly with respect to data and the modelling of HCE. Various models may yield different results, which should be taken into account when reading the results.

2.1 Review of micro-level studies

Micro-level studies allow for the inclusion of demographic and socio-demographic variables in particular. One focus in micro-level studies has been to test whether health care costs are explained by age per se and/or rather by the proximity to death. Obviously, the effect of population ageing cannot be estimated from a cross-sectional analysis within a given country, but the application of recurrent cross-sectional data over time may give some indication thereof.

Some studies concluded that there was an insignificant or no effect of ageing per se on health care costs per capita (Lubitz and Riley, 1993; Zweifel et al., 1999; Felder et al., 2000; Hogan et al., 2001; Spillman and Lubitz, 2000; Hoover et al., 2002; Werblow et al., 2007; Felder et al., 2008) while others did find an effect of ageing (Roos and Roos, 1987; Lubitz et al., 1995; Seshamani and Gray, 2004a, 2004b). Thus, Seshamani and Gray (2004a) used an English dataset from 1970 to 1999 of the cost of hospital treatment for patients who were 65 years and above in 1970. They found a significant effect of ageing in one model with the cost for hospital treatment increasing until the age of 80 or 85, but thereafter decreasing due to a declining probability of being hospitalised for the oldest old. Seshamani and Gray (2004b) used the same data set and found an elasticity of ageing (population above 65) of 0.027, implying that a 10 per cent increase in the population share above the age of 65 would imply an increase in the costs of hospital treatment of 0.27 per cent. Similarly, Stearns and Norton (2004) used Medicare data from 1992 to 1998 of individuals aged between 66 and 99 years. They used a two-part model and found positive effects for all age groups except for the age group 95-99 years. Further, Dormont et al. (2006) found that the rise in HCE due to population ageing in France has been relatively small and that decreased morbidity has induced savings, which has more than offset the increase in spending due to ageing.

Other studies, especially some which aimed at projecting future expenditure based on cross-sectional observations and population forecasts,

predicted increasing HCE due to ageing per se (Serup-Hansen et al., 2002; Schulz et al., 2004). The latter study used German data and controlled for proximity to death and concluded that population ageing would cause a moderate increase in hospital days, but a much stronger demand for long-term care. Moreover, it was found that changes in the disease patterns would require a reallocation of hospital resources.

Quite a large number of studies have found increasing health care costs on acute health care by increasing the proximity to death (Roos and Roos, 1987; Lubitz and Riley, 1993; Lubitz et al., 1995; Spillmann and Lubitz, 2000; Zweifel et al., 1999; Felder et al., 2000; Hogan et al., 2001; Hoover et al., 2002, Seshamani and Gray, 2004a, 2004b; Serup-Hansen et al., 2002; Stearns and Norton, 2004; Schulz et al., 2004). In a later study, Breyer and Felder (2006) confirmed an effect of adjusting for “costs of dying” in a projection of future costs. One study found no association between the proximity to death and the use of GP services (Madsen et al., 2002), while another found only a minor association between the proximity to death and the use of prescribed drugs (Kildemoes et al., 2006). In contrast, the cost of long-term care as a function of the proximity to death has only been sparsely studied. Karlsson et al. (2006) found that the demand for formal long-term care will take off around 2015 and will reach a peak at some point after 2040. However, the most significant increase will be demand for informal care. Few studies found that the acute care cost of dying decreased above a certain age (Lubitz et al., 1995; Serup-Hansen et al., 2002; Seshamani and Gray, 2004a, 2004b) which reflects the decreasing probability for the oldest old of being hospitalised and/or the reduced probability of initiating intensive treatment.

In a Danish study, Arnberg and Bjørner (2010) estimated the age-specific public spending on health on basis of a 10 per cent random sample of the Danish population from 2000 to 2007 and register data on the utilisation of health care. In their study, they separated the effect of age and the effect of proximity to death. The results were combined with a long-term population forecast, and they predicted the effect of demographic change (both cohort differences and increased life expectancy) on public health expenditure. They found that while a constant life expectancy would increase public health expenditure by 17 per cent in 2050, due to a cohort effect, increased life expectancy would increase public health

expenditure by yet another 10 per cent after adjusting age-specific expenditure for proximity to death.

Breyer and Felder (2006) used 1999 Swiss data to derive age-specific costs and applied these to a population projection for Germany, restricted to the age interval of 30-95 years. The data allowed an estimation of the effect of both age and time until death. They estimated a purely demographic HCE growth by 0.37 per cent per year from 2002 to 2040, which corresponds to a total increase of 19.5 per cent. When taking the proximity to death into consideration, the growth forecast is reduced by one fifth.

A Finnish study (Häkkinen et al., 2008) was based on a sample of 40 per cent of the Finnish population aged 65 and above at the end of 1997 and projected expenditure on long-term care, somatic specialised care, health centre and psychiatric inpatient care and prescribed medicine until 2036. Their model included age, proximity to death and income. They concluded that although total HCE and care of elderly will increase by age, the pattern will be different for different categories of care. They conjectured that HCE is driven more by the use of long-term care and medical technology than age and gender alone, and therefore policy actions may be the main determinant of HCE.

A further development in the literature has been to look into the various factors that change with age and affect aggregate HCE differently at different age levels. Thus, Chernichovsky and Markowitz (2004) found that ageing and its correlates produce a complex picture of the potential effect of ageing on the total costs of medical care. Among such factors, they studied age-specific changes in morbidity and mortality, growth in income and insurance coverage, a rising level of education and changing technology. They concluded that shifting morbidity and mortality to older ages does not necessarily imply increasing HCE. Cutler et al. (2011) used data from a US Medicare Current Beneficiary Survey 1991-2007. They considered 19 different indicators of health in elderly and combined these into three broad summary categories, such as severe physical and social incapacity, less severe incapacities and light impairments related to vision and hearing. They found that the prevalence of the first and third category had declined rapidly over time, in contrast to the second category, and they concluded that the ageing process itself had changed over time. How this would impact health care expenditures was not investigated, though.

Results from studies on historical data are obviously dependent on both demand-side changes and possible supply-side reactions and, consequently, the observed HCE can be interpreted as utilisation rather than demand for health care. In contrast, the results from projections are dependent on the specific assumptions with respect to future age-specific health problems which have also been applied to the supply-side reaction. Most projections are based on an unchanged age-specific utilisation rate and unchanged health care rationing and can be interpreted as projections of demand.

2.2 Review of macro studies

With respect to macro studies, the determinants of HCE are an area that requires a specific understanding of the mechanisms that determine HCE at the national level. Especially, the budgets are to a great extent determined in a political process. Most studies that have been carried out so far have been based on a weak theoretical foundation (Roberts, 1999; Gerdtham and Jönsson, 2000). The choice of model as well as the explanatory variables appear more or less atheoretical, and studies that have actually used a theoretical foundation have apparently not succeeded in doing so satisfactorily. In addition, the quality of data has been far from perfect (Gerdtham and Jönsson, 2000), as they have been generated from different countries with different principles for national accounting and definitions of HCE, despite the efforts by the OECD to produce commensurable data (OECD, 2005). In particular, the extent to which care for the elderly has been included varies between countries. One further problem is associated with describing the characteristics of each country unequivocally by a limited number of variables, as no country has based its health care system entirely on a single model. Rather, institutional characteristics are most often a mixture of different models. For instance, co-payment may be the general rule, but exceptions may exist for certain types of medicine or groups of patients, or global budgets may be the general rule, but some activity-based payment may exist alongside the global budget.

According to Gerdtham and Jönsson (2000), it is possible to distinguish between two generations of studies, which are briefly reviewed below. Most studies are limited to OECD countries. The first generation

of studies can be characterised by being based on cross-sectional data alone. In his pioneering article, Newhouse (1977) carried out a bivariate regression of HCE in 13 countries, based on 1971 data. He found that income (GDP per capita) accounted for 92 per cent of the variation in HCE per capita, only leaving little room for other explanatory variables. Leu (1986) took the analysis one step further by including a multivariate regression with economic, demographic and institutional variables that were founded in public choice theory. Thus, besides income, he used age composition, resource variables and system variables to characterise each country's health care system. In a pooled cross-section time series analysis, Gerdtham et al. (1992b) used GDP, inflation rate, percentage of public HCE to total HCE and age composition, which explained between 83 and 98 per cent of the variation.

The second generation of analyses is characterised using panel data that allow more observations to be included (among others, Gerdtham et al., 1992b, 1998; Hitiris and Posnett, 1992). This also allows for dynamic aspects to be included, and for the inclusion of dummy variables to express country as well as time-specific effects that allow for unobservable variables which are correlated with the explanatory variables. The latest development uses specific methods to test for stationarity of data, i.e. whether there are genuine causal effects of the explanatory variables on HCE or whether the effects are spurious and caused by common time trends, since it has been demonstrated that non-stationarity can produce spurious results (Philips, 1986; Engle and Granger, 1987). Roberts (1999) addressed, in particular, issues related to dynamics and heterogeneity of data, and lack of sensitivity testing. She dealt with these three issues by adopting recently developed techniques for analysing dynamic heterogeneous data fields containing non-stationary variables.

While later studies have gradually refined the econometric methods, the main conclusions from previous studies still seem valid. A number of studies confirm that GDP per capita is the most important determinant of HCE in OECD countries, and that the effect of demographic and institutional variables is limited, yet measurable. A few results of age composition as well as economic, institutional and technology variables are reviewed in the following sections.

2.3 Age as a determinant of health care expenditure in macro studies

Age composition as a determinant of HCE was absent from the often quoted work by Newhouse (1977), but has been included in most macro studies since then. In his study of HCE, Leu (1986) included age composition and found that the share of the population below 15 years, but not above 65, was positively associated with HCE per capita. The last finding was unexpected. A later study by Getzen (1992) used OECD data from 20 countries in the period 1960-1988 and performed various cross-section analyses. He found no effect of age and claimed this finding to be consistent with the hypothesis that need, as measured by changing age composition, will affect the allocation of HCE between age groups in a population rather than the total budget. The hypothesis rests on the assumption that total health care spending is a result of political and institutional choices, rather than trends in demography, morbidity or technology. As opposed to these, Hitiris and Posnett (1992) used data from 20 OECD countries for the period 1960-1987. They regressed HCE as a function of GDP per capita, the share of the population above 65 years and the public finance share of total HCE. The elasticity of HCE with respect to the proportion of the population over 65 was around 0.55 and significant. This is a relatively high estimate, but the authors do not discuss this result. To a certain extent in accordance with the latter result, Gerdtham et al. (1992a) used data from 19 OECD countries for 1987 and replicated Leu's study. In one regression specification, they found a significant, but small positive effect of ageing. In other specifications, the estimates were insignificant and close to zero. In the same year, Gerdtham et al. (1992b) used data from 19 OECD countries for the years 1994, 1987 and 1980 in a pooled cross-section regression analysis and used economic, demographic and institutional variables. The age variable, measured by the ratio of 65+ year-olds/15-64 year-olds, showed a measurable positive effect with an elasticity of about 0.2. In a later panel data analysis, Gerdtham et al. (1998) applied data from 22 OECD countries for the period 1970-1991 to examine the effect of a number of institutional and non-institutional variables on HCE. Here, age composition was not found to have any significant effect. In all model specifications, the elasticity of HCE with respect to the proportion above 75 years was negative, but close to zero. These results were supported by Barros (1998) who used data from 24 OECD countries for the period 1960-1990 and estimated

growth rates rather than absolute levels of HCE. The explanatory variables included economic, demographic and institutional variables. One conclusion was that ageing has not contributed to the growth of HCE. Likewise, Roberts (1999) did a study based on data from 20 OECD countries from 1960 to 1993 and included the following variables: GDP per capita, public spending as a percentage of the total, the percentage of the population aged over 65 and the relative price of health care. She found a negative, but insignificant effect of ageing in alternative approaches. The negative sign was unexpected. However, in a panel approach, Hitiris (1997) used OECD data and studied determinants of HCE in 10 EC member states from 1960 to 1991 and found a positive association between the dependency rate (the population aged 0-19 plus 65 and over as a share of the population between 20 and 64) and HCE. In a re-examination of this study, Roberts (2000) found a positive association between the population share above 65 and HCE with a significant elasticity of 0.049. A 10 per cent increase of the population share above 65 would thus imply an increase in HCE by 0.49 per cent.

A further number of macro analyses are reviewed in Gerdtham and Jönsson (2000). They conclude that the effect of age structure is usually insignificant. However, Di Matteo and Di Matteo (1998) used Canadian provincial data from 1965 to 1991 to estimate the determinants of HCE and found that the proportion of the provincial population aged above 65 had an elasticity of 0.81. On basis of this, they estimated an annual increase of HCE by 1.3 per cent due to ageing for the period 1991-2025. Di Matteo (2005) also used American state-level data for the period 1980-1998 and Canadian province level data for the period 1975-2000 to examine the determinants of HCE per capita. In a model where time as a proxy for technology changes was included, ageing explained 8.9 per cent of the growth of HCE in the US and 10.3 per cent in Canada during the periods studied. In comparison, time/technology explained 62.3 per cent and 64.2 per cent, respectively.

Giannoni and Hitiris (2002) analysed determinants of regional HCE in Italy and included ageing (percentage of the population above 65) as a variable. They found an elasticity of 0.16 for this variable. Thus, a 10 per cent increase in this share would imply an increase in HCE by 1.6 per cent. Likewise, Crivelli et al. (2006) found a positive association between HCE and the proportion of the population above 75 in 26 cantons of

Switzerland with elasticities between 0.22 and 0.35. Further, Karatzas (2000) used OECD, IMF and UN databases to estimate the impact on US HCE of various determinants during the previous three decades. In an equation with income and ageing (population above 65) as the only determinants, he found a strong effect of ageing with an elasticity of 2.6 for total HCE and 3.3 for public HCE which – compared to most other studies – are extreme results. Later, Mosca (2007) studied the main determinants of HCE in 20 OECD countries during the period 1990 to 2000. Demography was included for the proportion of the population above 80 years resulting in an elasticity of 0.14.

Opposed to these studies, Herwartz and Theilen (2003) used OECD data from 1960 to 1997 to study variations in HCE relative to income, ageing (population above 65) and technological change. They found that the association with ageing was small and insignificant. They also found an indication of a divergence in cost-containment policies across countries. Likewise, in his essay on ageing and health care costs, Reinhardt (2003) draws on research literature and concludes that ageing is nowhere the strongest driver of demand for health care in the US. Dominant drivers in the past as well as in the future are increases in per capita income, costly medical technology, increasing workforce costs due to shortage and an asymmetric distribution of market power in health care in favour of the supply side. One may argue, however, that this relates to the US, but not necessarily to Europe, which has a lower level of HCE, stronger political regulation and different institutional structures compared to the US. Reinhardt further points out that the prevailing age-specific health care utilisation pattern can be changed, and that age-specific spending does not necessarily have to increase for all age groups in the future.

2.4 Technology

The question of how HCE is affected by technology change seems highly relevant. In particular, one may expect that technology development may benefit the elderly to a larger extent compared to the younger due to higher morbidity among the elderly and, therefore, technology can be expected to be used to a greater extent to treat elderly patients. The question has been considered or explicitly analysed in a few studies, but – as pointed out by Dybczak and Przywara (2010) – there are many factors

which affect HCE, and due to the complexity of their interaction, it is difficult to identify the individual effect of each of these on HCE. They claim that there are no reliable forecasts of how technological development impacts on future HCE. “Technology” has often been subsumed into a residual factor that covers factors not explicitly accounted for in the analyses. Even time has been used as a proxy for technological development as in the study by Di Matteo (2005).

In his study, Newhouse (1992) pointed to technological changes as a factor that accounts for about 75 per cent of the increase in HCE during the previous 50 years. Okunade and Murthy (2002) followed up on this study in an econometric analysis using time series data of US health care costs for 1960-1997 and used R&D expenditure as proxies for technology changes. They found that technology change was a statistically significant long-run driver of HCE.

In a comprehensive study covering 20 countries with data from 1981 to 2002 (OECD 2006), various ageing scenarios were applied. Under the assumption of “healthy ageing”, it was found that longevity only has a modest effect on HCE. It was estimated that between 1981 and 2002, total HCE on average increased by 3.6 per cent per year, where only 0.3 percentage points were accounted for by demographic effects and 2.1 percentage points by income effects, while the rest (about 1 percentage point) was referred to as a residual effect, covering technology and relative prices. The results are complicated, though, due to cost-containment policies in some countries like Denmark and Sweden that have resulted in low residuals.

Dybczak and Przywara (2010) based their empirical study on 20 OECD members from Europe, and they used two variables to represent demographic development (the share of the population below 20 years and the share above 80 years). Their main conclusion was that the effect of demographic change is relatively weak compared to the effect of non-demographic factors (including technology change) over the last decades.

2.5 Other determinants of health care expenditure in macro studies

Other determinants that have been used are variables that can be classified as economic, social, behavioural or institutional, although a distinction between these types is not clear-cut. From a public choice perspec-

tive, Leu (1986) included, among other variables, public provision and financing as a share of total HCE and he expected a positive association with HCE. These hypotheses have been challenged, however, by, among others, Culyer (1989). Further economic and institutional variables that have been used to describe a health care system are gate-keeping by GPs, the number of doctors per citizen and the number of beds. Case-mix has been accounted for by the ratio of in-patient spending to total spending. Female labour force participation has been seen as an indicator of substitution of formal care for informal care, leading to higher HCE (Gerdtham et al., 1992a, 1992b).

Among the institutional variables are variables that catch the incentives built into the payment to providers, or variables that regulate total expenditure. Thus, fee-for-service versus capitation payment of GPs or payment of salaries, open-ended versus close-ended budgets (or global budgets) of hospital payment and budget ceilings have been used (Gerdtham et al., 1992b, 1998). On the demand side, co-payment is assumed to affect utilization. In the framework by Hurst (1991) and later OECD publications (1992, 1994), health care systems are classified as either public integrated systems with universal coverage by one third-party payer and mainly public provision of health care, public contract systems with universal coverage by multiple insurers and contracts with public or private providers, or a mix of these. Urbanisation has been included as an indicator of travel costs of health care utilisation. Finally, the unemployment rate has been used in some studies (e.g. Gerdtham and Jönsson, 2000).

In a review, Gerdtham and Jönsson (2000) summarised the existing empirical results with a specific reliance on the comprehensive study by Gerdtham et al. (1998) as follows: a variable that consistently shows a positive association with HCE is GDP/capita, while the age structure of the population, the unemployment rate and female labour force participation are usually insignificant. Six results from their 1998 study seemed reasonably strong and corresponded to the expectation: Hence, HCE seems lower where 1) primary care gatekeepers are used, 2) patients pay the provider and get reimbursed afterwards, 3) capitation systems are used as compared to the use of fee-for-service systems, 4) there is a high reliance on outpatient visits as compared to inpatient care and 5) there is public sector provision of health care (although with some reservations),

while 6) the total supply of doctors may be positively associated with HCE. Other results did not correspond to the expectations: budget ceilings on inpatient care seemed to be unexpectedly associated with a higher HCE, while systems with public reimbursement tended to have a lower HCE than contract systems, and public integrated systems may be more costly than public contract systems. The authors assume that the last finding may be explained by countries in the public integrated system also tending to "... have higher fractions of high-cost in-patient care and fewer gate-keeping arrangements" (p. 47). They found it difficult, though, to explain the result for public reimbursement systems.

The overview reveals that the majority of the macro studies apply OECD data, which obviously has the advantage that data are easily available, and the OECD has made a large effort to provide consistent measures. Furthermore, there seems to be a considerable variation in the health care system characteristics and in the other explanatory variables across OECD countries which are important in the regression analysis. The review also shows that no clear-cut conclusion can be inferred with respect to the effect of ageing on HCE, although most studies point towards modest increases in HCE due to ageing societies.

3. Our empirical study

Ageing and aggregate HCE in EU-15 countries were studied by the present authors and colleagues (Christiansen et al., 2006; Bech et al., 2011) and the results are summarised below.

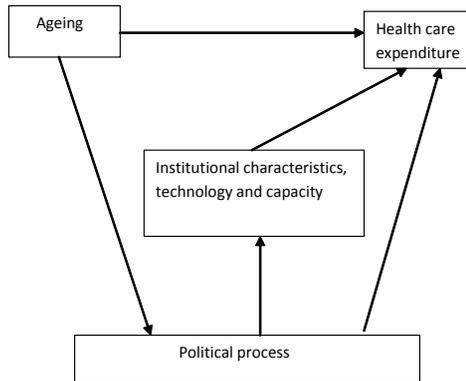
As described in Section 1 above, various models for longevity and health at the end of life have been formulated. In the following sections, we assume as a working hypothesis that ageing increases demand per capita.

3.1 Hypotheses

Our basic thinking was as follows and as illustrated in Figure 3: we assumed that ageing increases the demand for health care, and that this demand leads both directly to a change in HCE (within a given institutional setting and a given capacity) and indirectly to changed expenditure

through changes in institutional and capacity variables. The changes in these variables are assumed to take place in a political process, which may mediate the direct influence of ageing. HCE may also increase directly through a political decision. The political process is considered as a “black box” in the present study.

Figure 3. Model of health care expenditure



Source: Authors' illustration.

Technological changes may affect the cost of care for all age groups, but eventually various age groups will be affected differently. It is an empirical question whether improved technology will affect the older age groups more than the rest. Consequently, the indicator of technology is assumed to increase the level of spending in general.

While the principal aim was to study the effect of ageing populations on HCE, other variables with an assumed effect on HCE have also been included, such as economic variables (HCE per capita); social variables (female labour force participation; unemployment); behavioural variables (alcohol and tobacco consumption); structural variables (public integrated or contract system, free choice of hospital, gate-keeping by GPs); mode of financing providers, copayment, and hospital capacity (beds, physicians) and technology (dialysis and scanning capacity).

3.2 Data and model

The data used in the study were an unbalanced panel data set that covers the old 15 European Union member countries (EU-15), i.e. Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden and United Kingdom. The panel spans a time period of 24 years (1980-2003).

The dependent variable was total health care expenditure per capita (THEPC) measured in US dollars in nominal prices, adjusted for purchasing power parities (PPP) and for inflation.

The explanatory variables and their expected association with the dependent variable are listed in Table 1. These variables have been collected from a variety of sources including OECD Health Data (OECD, 2004), WHO (2004) and EUROSTAT (European Commission, 2004). Since the study is a macro-level study, it is not relevant to include proximity to death as a variable.

A sequence of models with THEPC as the dependent variable was estimated resulting in a full model which included all variables as well as the country effects and a time trend. For any regression, THEPC and GDP were log transformed prior to the regression in order to account for potential right skewness of these variables.

Table 1. Explanatory variables

	Name	Description	Data source	Hypothesised Influence
Economic, social and demographic variables	GDP	Gross domestic product per capita, USD in nominal prices and adjusted for PPP	OECD/WHO/Eurostat	+
	AGE 0-5	Proportion of the population aged 0-5 (%)	EUROSTAT	+
	AGE 65-74	Proportion of the population aged 65-74 (%)	EUROSTAT	+
	AGE 75+	Proportion of the population aged 75-84 (%)	EUROSTAT	+
	ALCCON	Alcohol consumption (litres of pure alcohol per capita)	OECD/WHO	+
	TOBCON	Tobacco consumption (cigarettes per capita)	OECD/WHO	+
	FLFPR	Female labour force participation rate (% ratio to active population aged 15-65)	OECD	+
	UNEMP	Unemployment rate (% ratio to labour force)	ILO	+
	LE 65 F	Life expectancy at age 65 for females	WHO	+
	LE 65 M	Life expectancy at age 65 for males	WHO	+

Table 1. Continued....

	Name	Description	Data source	Hypothesised Influence
Institutional variables	PUBCONTR	Dummy variable for countries characterised as a public contract health care system, zero otherwise	Own	+
	PUBINT	Dummy variable for countries characterised as a public integrated health care system, zero otherwise	Own	+
	PUHES	Public health expenditure as share of total health expenditure	OECD/WHO	+
	GATE	Dummy variable for countries with physicians as gatekeepers, zero otherwise	Own	+
	COPAYGP	Dummy variable for countries with significant co-payment for GP, zero otherwise	Own	+
	COPAYHO	Dummy variable for countries with significant co-payment for inpatient hospital treatment, zero otherwise	Own	+
	GLOBALHO	Dummy variable for countries with global budget reimbursement of hospitals, zero otherwise	Own	+
	CASEHO	Dummy variable for countries with case-based reimbursement of hospitals, zero otherwise	Own	-
	SALARYGP	Dummy variable for countries with salaried GPs, zero otherwise	Own	+
	CAPGP	Dummy variable for countries with capitation payment of GPs, zero otherwise	Own	+
	CEILHO	Dummy variable for countries with overall ceiling on hospitals, zero otherwise	Own	+
	FREEHO	Dummy variable for countries with free choice of hospital, zero otherwise	Own	+
	FREEGP	Dummy variable for countries with free choice of GP or primary care physician, zero otherwise	Own	+
Tech. and capacity variables	PHYS	Physicians per 1 000 inhabitants	OECD/WHO	+
	BEDS	Acute care beds per 1 000 inhabitants	OECD/WHO	+
	DIALY	Patients undergoing dialysis per 100 000 inhabitants	OECD	+
	MRIU	Magnetic resonance imaging units per 1 million inhabitants	OECD	+

Source: See table.

3.3 Results

Table 2 shows regression results from the stepwise build-up. Country-specific dummies as well as time trend have been included in models (3)-(6). The final model (6) shows a slight improvement over models (3) and (4) as measured by the adjusted R-square. Apparently, much variation has been included in the country-specific dummy variables and the time trend. Thus, it is seen that R^2 increases substantially when these are included (compare models 1 and 3). This is to be expected, as the country dummies account for any (time invariant) variation in THEPC across countries, to the extent that this variation is not captured by the explanatory variables. This includes in particular the effects of time invariant institutional variables. We considered estimating the model without country dummies in order to investigate these institutional effects. However, this was not a giving exercise, as the effects from the time invariant institutional characteristics could not be separated from country-specific effects in such a simplified setup.

Furthermore, a time trend is included to account for a (country invariant) increase in THEPC. The results not shown here indicated a significant increase over time, which increased the R^2 by approximately 8 percentage points. Although the included supply and demand variables do play a role, the high significances and R^2 values caused by the inclusion of the country dummies and the time trend demonstrate that there are many causes for variation in the THEPC which cannot be explained by the observed explanatory variables.

While model (1) includes age composition, model (2) includes life expectancy as explanatory variables. The latter variables are included as they indirectly relate to age composition. As age composition and life expectancy are clearly correlated, models (5) and (6) are analysed without and with life expectancy to identify the effect of age composition alone (model 5). The coefficients in the table should be interpreted as elasticities. For variables which are measured as percentages (the age group variables, FLFPR and UNEMPL) the coefficient should be interpreted as semi-elasticities; that is, the percentage increase in THEPC per capita associated with a 1 percentage point increase in the variable. It appears that some of the coefficients are rather small. Thus, a 10 percentage point

increase in the age group 65-74 years is associated with an increase of 0.3 per cent in expenditure in EU-15 (cf. model 4).

Table 2 provides mixed evidence as regards the impact of age structures for EU-15. Model (1) indicates a negative effect in the age group 0-5 years.

In the age group 65-74 years, the simple model (1) indicates a negative association. However, when country heterogeneity (models 3-6 include country dummies) is accounted for, the effect turns positive. This is an important observation, as it seems to indicate a potential cross-country spurious relation between age 65-74 and THEPC. That is, countries with high proportions of 65-74-year-olds have low THEPC for reasons unrelated to age. The effect remains positive when we account for GDP and population characteristics (model 4), but turns out to lose significance when we account for institutional and technology variables (model 5). Presumably, some of the age-related effects are subsumed in the institutional variables as hypothesised in our model. In particular, it can be observed that indicators of use of technology (DIALY, MIRU) and capacity (BEDS) have positive statistical signs. Still, PHYS as a capacity variable is weakly negative.

Considering the age group 75+ years, the simple model (1) reports a positive effect. This changes to significantly negative when country dummies and the time trend of THEPC is accounted for (model 3). This is also an important observation, as it presumably indicates the presence of a spurious correlation over time between age 75+ and THEPC: both have been increasing throughout the period, so that an unconditional positive relationship seems to exist. This negative effect is robust towards the inclusion of GDP and population characteristics (model 4), but turns insignificant when we control for institutional and technology variables. In return, the life expectancy variables for males and females at 65 turn significant, although with smaller numerical effects compared to model (2).

Table 2. Regressions (N=15, T=24, N*T=360). Dependent variable: log THEPC

	(1)	(2)	(3)	(4)	(5)	(6)
INTERCEPT	7.09 (0.23)***	4.91 (0.20)***	7.65 (0.15)***	0.35 (0.76)	-1.03 (0.06)**	-0.51 (0.55)
AGE 0-5	-0.049 (0.014)***		-0.045 (0.008)***	-0.013 (0.007)*	-0.009 (0.006)	-0.007 (0.005)
AGE 65-74	-0.059 (0.018)***		0.017 (0.008)**	0.030 (0.008)***	0.015 (0.006)***	0.007 (0.005)
AGE 75+	0.192 (0.016)***		-0.059 (0.016)***	-0.050 (0.013)***	-0.006 (0.010)*	-0.011 (0.010)
LOG GDP				0.714 (0.087)***	0.693 (0.060)***	0.559 (0.061)***
LE 65 F		0.203 (0.022)***		0.021 (0.020)	-	0.067 (0.014)***
LE 65 M		-0.086 (0.025)***		-0.038 (0.024)	-	-0.029 (0.016)*
FLFPR				0.001 (0.004)	0.048 (0.003)*	0.005 (0.002)*
UNEMP				-0.002 (0.002)	-0.002 (0.001)	-0.004 (0.001)***
ALCCON				-0.008 (0.006)	-0.009 (0.005)**	-0.001(0.004)
TOBCON				0.00006 (0.00002)***	0.00003 (0.00001)**	0.00004 (0.00001)***
CEILHO					-	-
PUHES					0.010 (0.001)***	0.009 (0.001)***
GATE					-	-
PUBINT					-	-
PUBCONTR					-	-
SALARYGP					-0.101 (0.039)***	-0.110 (0.037)***

Table 2. Continued...

	(1)	(2)	(3)	(4)	(5)	(6)
CAPGP					-0.096 (0.033)***	-0.102 (0.031)***
GLOBALHO					0.026 (0.013)*	0.006 (0.013)
CASEHO					-0.084 (0.028)***	-0.098 (0.026)***
COPAYGP					0.071 (0.038)*	0.061 (0.035)*
COPAYHO					-0.024 (0.020)	-0.021 (0.019)
FREEGP					-0.009 (0.038)	0.013 (0.036)
FREEHO					0.082 (0.046)*	0.045 (0.044)
PHYS					-0.011 (0.0004)***	-0.00099 (0.00039)**
DIALY					0.001 (0.0006)**	0.001 (0.0005)***
MRIU					0.004 (0.002)**	0.003 (0.001)**
BEDS					0.064 (0.007)***	0.077 (0.007)***
R^2 / R^2_{ADJ}	0.41 / 0.41	0.32/0.32	0.91 / 0.90	0.94 / 0.94		0.98 / 0.98
R^2 / R^2_{ADJ} (model with country dummies only)			0.62 / 0.61			
R^2 / R^2_{ADJ} (model with country dummies and time trend only)			0.89 / 0.89			

Source: Authors' calculations.

Note: Models (3) to (6) are adjusted for country effects and a time trend. Significance marked by ***(1%), **(5%), *(10%). The dependent variable is log (THEPC). CEILHO only varied for Austria and Belgium (from 1 to 0) and for Finland and Italy (from 0 to 1). For Denmark, Germany and the Netherlands it was constant at 0. For the remaining countries it was missing. GATE assumed a constant value 0 for Belgium, France, Germany, Greece and Luxembourg, and 1 for the remainder. PUBINT assumed a constant value 0 for Austria, Belgium, France, Germany, Greece, Luxembourg, Netherlands and Portugal, and 1 for the remainder. PUBCONTR assumed a constant value 1 for Austria, Belgium, Germany, Luxembourg and Netherlands, and 0 for the remainder.

Thus, to summarise, when we control properly for unconditional country heterogeneity and increasing HCE over time, there is an indication of a negative relationship between THEPC and the age groups 75+, while a positive relationship between HCE and age 65-74 seems to exist. These effects are robust toward control for GDP and population characteristics, but they seem to be related to institutional and technology characteristics, as they lose significance when these are controlled for.

Our data were further analysed by the use of a model with lagged values of 1 year of the dependent as well as the explanatory variables to get further insight. A number of statistical tests for the underlying assumptions were performed.

We found that HCE per capita was increasing with an increasing share of the population in the age groups 65-74 years and 75+ years. However, an almost equal, but negative, association was found for lagged values of the same age groups (size of age groups in the previous year). The implication was that a high proportion of elderly did not in itself drive HCE. In other words, we found a short-run positive effect of ageing on HCE, but in the long run the effect was almost zero. In return, life expectancy and mortality were positively associated with HCE as expected. Thus, the coefficient of life expectancy at 65 was -0.0087 , but insignificant, while the coefficient of the lagged value of the variable was 0.0352 and significant. The coefficient of mortality of men was -0.1070 and insignificant, while the coefficient of the lagged value was 0.1689 and significant. It is interesting to notice that it was the lagged values rather than the present values that were associated with increased HCE. The interpretation is that it takes some time from a shift in these variables to policy implementation. The combined effect of changing age groups and changed life expectancy could not be derived, however, since the results were partial effects.

3.4 Discussion of our empirical results

It is worth noticing that the empirical findings are based on historical data and are a result of former demand pressure and deliberate supply-side responses and/or built-in mechanisms in health care systems in Europe. However, with this reservation in mind, several relevant points may be raised for discussion. The estimated age effects seem to be mediated by

the institutional and technological characteristics, as they turn insignificant when we control for these. Whether such a mediating effect has actually occurred, or whether there are other reasons why institutional variables explain much of the variation which ageing accounts for in a simpler model, cannot be tested in the present study. More fundamentally, we do not have information about the policy processes where adjustments in health care services are made in response to changes in demography.

At a more detailed level, it can be concluded that an increase in the age group 65-74 as a share of the population is positively associated with HCE per capita as expected. This effect turns insignificant when institutional and technological characteristics are included. For the age group 75+, the association with health expenditure turned out to be negative but insignificant when institutional and technology variables are included. Thus, it seems to be the case that it is the presence of specific institutional structures and health care technology which benefits very old people, rather than that it is the proportion of such people which governs HCE.

4. Summary and discussion

In summary, a review of the literature shows different conceivable scenarios of ageing with different effects on the projected future HCE. Thus, an extension of the morbidity scenario implies increased HCE, while a compression of the morbidity scenario will entail reduced HCE. A modified model with a later onset of morbidity combined with an increased length of life will eventually entail moderate increases in HCE.

These scenarios assume that technology is unchanged. We found no model of how an increased use of advanced technology might affect future HCE. One may speculate, though, that advances in technology will add to the HCE of the elderly, in particular due to higher morbidity among the elderly.

The empirical literature clearly shows that income is a main driver of HCE, although HCE is publicly regulated in most countries. This indicates that what a country spends in general depends on “what it can afford” and how it prioritises. Hence, ageing cannot be expected to automatically increase HCE, even though ageing would entail increased demand.

This leaves us with a model where the consequences of ageing on HCE may manifest themselves either directly through increased demand within a given capacity, or through a policy process where the capacity or institutional set-up changes in response to demographic changes.

The empirical literature comprises two types of studies, namely micro-level and macro-level studies. Micro-level studies are based on the observation of individuals while macro-level studies are based on countries or regions as the unit of observation. It has been observed in many micro-level studies that health care costs increase by increased proximity to death. As life expectancy increases, the costs will gradually occur later in time which has to be taken into account when prognoses are made of future HCE. Forecast results are highly dependent on the assumptions with regard to demographic changes and age-specific utilisation. It has been conjectured that HCE will be driven more by long-term care and medical technology than by demographic changes, and that policy actions may be a main determinant of HCE.

In contrast to micro-level studies, macro-level studies allow for the inclusion of institutional and population characteristics along with income in explaining HCE. Most studies conclude that ageing only has a relatively small effect on future HCE, as HCE is explained by income and other variables.

Moïse and Jacobzone (2003) concluded that HCE for the very old is lower than for younger age groups because, on average, they receive less aggressive and less expensive treatments. Their conclusion, however, was less clear-cut when long-term care was included. A practice style discriminating the older age groups may to some extent be in play. In future studies, it might be worthwhile to follow up on the study of age-related diseases by Moon et al. (2003) and look into diagnoses to see whether there is an age dependency in the incidence and costs of treating specific health problems. The hypothesis would be that the incidence of health problems with high treatment costs varies by age. Or there may be an age variation in presenting health problems to professional health care givers as well as in the aggressiveness of treating presented problems. Obviously, this demand side consideration has to be connected to supply-side restrictions and prioritisations.

Our study has explicitly focussed on the effect of ageing on HCE, and the importance of allowing for proximity-to-death when forecasting HCE

from cross-sectional data was pointed out. However, as has become clear from the literature review, several factors besides demography contribute to HCE, not to mention policy processes.

References

- Arnberg, S. and Bjørner, T.B. (2010), Estimation af sundhedsudgifternes afhængighed af alder og afstand til død, DØR Working Paper 2010:1, Copenhagen.
- Bech, M., Christiansen, T., Khoman, E., Lauridsen, J.T. and Weale, M. (2011), Ageing and health care expenditure in EU-15, *European Journal of Health Economics* 12, 469-478.
- Barros, P.P. (1998), The black box of health care expenditure growth determinants, *Health Economics* 7, 533-544.
- Breyer, F. and Felder, S. (2006), Life expectancy and health care expenditures: A new calculation for Germany using costs of dying, *Health Policy* 75, 178-186.
- Chernichovsky, D. and Markowitz, S. (2004), Aging and aggregate costs of medical care: Conceptual and policy issues, *Health Economics* 13, 543-562.
- Christiansen, T., Bech, M., Lauridsen, J. and Nielsen, P. (2006), Demographic changes and aggregate health-care expenditure in Europe, ENEPRI Research Report 32, Brussels, <http://www.enepri.org/files/AHEAD/Reports/WP6-b.pdf>.
- Crivelli, L., Filippini, M. and Mosca, I. (2006), Federalism and regional health care expenditures: An empirical analysis for the Swiss cantons, *Health Economics* 15, 535-541.
- Culyer, A.J. (1989), Cost containment in Europe, *Health Care Financing Review (Ann Suppl)*, 21-32.
- Cutler, D.M. and Landrum, M.B. (2011), Dimensions of health in the elderly population, NBER Working Paper 17148.
- Di Matteo, L. (2005), The macro determinants of health expenditures in the United States and Canada: Assessing the impact, *Health Policy* 71, 23-42.
- Di Matteo, L. and Di Matteo, R. (1998), Evidence on the determinants of Canadian provincial government health expenditures: 1965-1991, *Journal of Health Economics* 17, 211-228.
- Dormont, B., Grignon, M. and Huber, H. (2006), Health expenditure growth: Reassessing the threat of ageing, *Health Economics* 15, 947-963.
- Dybczak, K. and Przywara, B. (2010), The role of technology in health care expenditure in EU, European Commission, Economic Papers 400, http://ec.europa.eu/economy_finance/publications/economic_paper/2010/pdf/ecp400_en.pdf.
- Engle, R.F. and Granger, C.W.J. (1987), Cointegration and error correction: Representation, estimation and testing, *Econometrica* 35, 251-276.
- European Commission (2004), EUROSTAT – Statistical Yearbook 2004, European Commission, Brussels, http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-AF-04-001/EN/KS-AF-04-001-EN.PDF.
- Felder, S., Werblow, A. and Zweifel, P. (2008), Do red herrings swim in circles?, *Ruhr Economic Papers* 73, Universität Duisburg, Essen.

- Felder, S., Meier, M. and Schmitt, H. (2000), Health care expenditure in the last month of life, *Journal of Health Economics* 19, 679-695.
- Fries, J. (1980), Aging, natural death, and the compression of morbidity, *New England Journal of Medicine* 303, 130-335.
- Gerdtham, U.-G. and Jönsson, B. (2000), International Comparisons of Health Expenditure: Theory, Data and Econometric Analysis, in A.J. Culyer and P. Newhouse (eds.), *Handbook of Health Economics Vol 1A*, Elsevier, Amsterdam.
- Gerdtham, U.-G., Sögaard, J., Andersson, F. and Jönsson, B. (1992a), Econometric analysis of health expenditure: A cross-sectional study of the OECD countries, *Journal of Health Economics* 11, 63-84.
- Gerdtham, U.-G., Sögaard, J., Jönsson, B and Andersson, F. (1992b), A pooled cross-section analysis of the health expenditure of the OECD countries, in P. Zweifel and H. Frech (eds.), *Health Economics Worldwide*. Kluwer Academic Publishers, Dordrecht.
- Gerdtham, U.-G., Jönsson, B., MacFarland, M. and Oxley, H. (1998), The determinants of health expenditure in the OECD countries, in P. Zweifel (ed.), *Health, the Medical Profession, and Regulation*, Kluwer Academic Publishers, Dordrecht.
- Getzen, T.E. (1992), Population aging and the growth of health expenditures, *Journal of Gerontology, Social Sciences* 47, S98-104.
- Giannoni, M. and Hitiris, T. (2002), The regional impact of health care expenditure: The case of Italy, *Applied Economics* 14, 1829-1836.
- Gruenberg, E.M. (1977), The failure of success, *Milbank Memorial Fund Quarterly, Health Soc.* 55, 3-24.
- Hagemann, R.P. and Nicoletti, G. (1989), Population ageing: Economic effects and some policy implications for financing public pensions, *OECD Economic Studies* 12, Paris, <http://www.oecd.org/dataoecd/17/18/35379092.pdf>.
- Herwartz, H. and Theilen, B. (2003), The determinants of health care expenditure: Testing pooling restrictions in small samples, *Health Economics* 12, 113-124.
- Hitiris, T. (1997), Health care expenditure and integration in the countries of the European Union, *Applied Economics* 29, 1-6.
- Hitiris, T. and Posnett, J. (1992), The determinants and effects of health expenditure in developed countries, *Journal of Health Economics* 11, 173-181.
- Hogan, C., Lunney, J., Gabel, J. and Lynn, J. (2001), Medical beneficiaries' costs of care in the last year of life, *Health Affairs* 20, 188-195.
- Hoover, D.R., Crystal, S., Kumar, R., Samamoohi, U. and Cantor, J.C. (2002), Medical expenditure during the last year of life: Findings from the 1992-1996, Medicare Current Beneficiary Survey, *Health Services Research* 37, 1625-1642.
- Hurst, J.W. (1991), Reforming health care in seven European nations, *Health Affairs* 10, 7-21.
- Häkkinen, U., Martikainen, P., Noro, A., Nithilä, E. and Peltola, M. (2008), Aging, health expenditure, proximity to death, and income in Finland, *Health Economics, Policy and Law* 3, 165-195.
- Karatzas, G. (2000), On the determination of the US aggregate health care expenditure, *Applied Economics* 32, 1085-1099.
- Karlsson, M., Mayhew, L., Plumb, R. and Rickayzen, B. (2006), Future costs for long-term care. Cost projections for long-term care for older people in the United Kingdom, *Health Policy* 75, 187-213.

- Kildemoes, H.W., Christiansen, T., Gyrd-Hansen, D., Kristiansen, I.S. and Andersen, M. (2006), The impact of population ageing on future Danish drug expenditure, *Health Policy* 75, 298-311.
- Kramer, M. (1980), The rising pandemic of mental disorders and associated chronic diseases and disabilities, *Acta Psychiatria Scandinavia* 62 (Suppl. 285), 382-397.
- Leu, R.E. (1986), The public-private mix and international health care costs, in A.J. Culyer and B. Jönsson (eds.), *Public and Private Health Services*, Basil Blackwell, Oxford.
- Lubitz, J.D., Beebe, J. and Baker, C. (1995), Longevity and Medicare expenditure, *New England Journal of Medicine* 332, 999-1003.
- Lubitz, J.D. and Riley, G.F. (1993), Trends in medicare payments in the last year of life, *New England Journal of Medicine* 328, 1092-1096.
- Madsen, J., Serup-Hansen, N., Kragstrup, J. and Kristiansen, I.S. (2002), Ageing may have limited impact on future costs of primary care providers, *Scandinavian Journal of Primary Care* 20, 169-173.
- Manton, K.G. (1982), Changing concepts of morbidity and mortality in the elderly population, *Milbank Memorial Fund Quarterly, Health Soc.* 60, 183-244.
- Michel, J.-P. and Robine, J.-M. (2004), A “new” general theory of population ageing, *Geneva Papers on Risk and Insurance* 29, 667-678.
- Moïse, P. and Jacobzone, S. (2003), Population ageing, health expenditure and treatment: An ARD perspective, in *A Disease-based Comparison of Health Systems. What is Best at What Cost?*, OECD, Paris.
- Moon, L., Moïse, P., Jacobzone, S. and the ARD-Stroke Experts Group (2003), Stroke care in OECD countries: A Comparison of treatment, costs and outcomes in 17 Countries, *OECD Health Working Papers* 5, <http://www.oecd.org/dataoecd/10/46/2957752.pdf>.
- Morrow, K.M. and Roeger, W. (1999), EU pension reform – An overview of the debate and empirical assessment of the main policy reform options, *Economic Papers* 162, The European Commission, Directorate-general for Economics and Financial Affairs, Brussels.
- Mosca, I. (2007), Decentralization as a determinant of health care expenditure: Empirical analysis for OECD countries, *Applied Economics Letters* 14, 511-515.
- Myrdal, G. (1940), *Population: A Problem of Democracy*, Harvard University Press, Cambridge.
- Myrdal, A. and Myrdal, G. (1934), *Kris i Befolkningsfrågan*, Nya Doxa, Stockholm.
- Newhouse, J.P. (1977), Medical care expenditure: A cross-national survey, *Journal of Human Resources* 12, 115-125.
- Newhouse, J.P. (1992), Medical care costs: How much welfare loss?, *Journal of Economic Perspectives* 6, 3-21.
- OECD (1992), *The reform of health care – A comparative analysis of seven OECD countries*, Health Policy Studies 2, OECD, Paris.
- OECD (1994), *The reform of health care systems – A review of seventeen OECD countries*, Health Policy Studies 5, OECD, Paris.
- OECD (2004), *Health data 2004*, OECD, Paris.
- OECD (2005), *Health update*, Internal Co-ordination Group for Health (ICGH), No. 1, OECD Paris, <http://www.oecd.org/dataoecd/10/57/35101765.pdf> (March 21 2012).

- OECD (2006), Projecting OECD health and long-term care expenditures: What are the main drivers?, Economics Department Working Papers 477, OECD, Paris, http://www.oecd.org/LongAbstract/0,3425,en_2649_34113_36085941_1_1_1_1,00.html.
- Okunade, A.A. and Murthy, N.R. (2002), Technology as a 'major driver' of health care costs: A cointegration analysis of the Newhouse conjecture, *Journal of Health Economics* 21, 147-159.
- Philips, P.C.B. (1986), Understanding spurious regressions in econometrics, *Journal of Econometrics* 33, 311-340.
- Reinhardt, U. (2003), Does the aging of the population really drive the demand for health care?, *Health Affairs* 22, 27-39.
- Roberts, J. (1999), Sensitivity of elasticity estimates for OECD health care spending: Analysis of a dynamic heterogeneous data field, *Health Economics* 8, 459-472.
- Roberts, J. (2000), Spurious regression problems in the determinants of health care expenditure: A comment to Hitiris (1997), *Applied Economics Letters* 7, 279-283.
- Roos, N.P. and Roos, L.L. (1987), Health care utilization in the years prior to death, *The Milbank Quarterly* 65, 231-254.
- Schulz, E., Leidl, R. and König, H.H. (2004), The impact of ageing on hospital care and long-term care – The example of Germany, *Health Policy* 67, 57-74.
- Serup-Hansen, N., Wickstrøm, J. and Kristiansen, I.S. (2002), Future health care costs – Do health care costs during the last year of life matter?, *Health Policy* 62, 161-172.
- Seshamani, M. and Gray, A.M. (2004a), Ageing and health care expenditure: The red herring argument revisited, *Health Economics* 13, 303-314.
- Seshamani, M. and Gray, A.M. (2004b), A longitudinal study of the effects of age and time to death on hospital costs, *Journal of Health Economics* 23, 217-235.
- Spillmann, B. and Lubitz, J. (2000), The effect of longevity on spending for acute and long-term care, *New England Journal of Medicine* 342, 1409-1415.
- Stearns, S.C. and Norton, E.C. (2004), Time to include time to death? The future of health care expenditure prediction, *Health Economics* 13, 315-327.
- UN (2011), World population prospects, Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, New York, <http://populationpyramid.net/>.
- Werblow, A., Felder, S. and Zweifel, P. (2007), Population ageing and health care expenditure: A school of 'red herrings'?, *Health Economics* 16, 1109-1126.
- WHO (2004), Tobacco-free Europe – Policy, WHO Regional Office for Europe, Copenhagen.
- Zweifel, P., Felder, S. and Meiers, M. (1999), Ageing of population and health care expenditure: A red herring?, *Health Economics* 8, 485-496.

Comment on Christiansen, Lauridsen and Bech: Ageing populations: More care or just later care?

Anna Lilja Gunnarsdottir *

The study is well done and interesting and contributes to the knowledge in this field. It is important to connect research and practice and to connect research and policy making.

Many OECD countries have been going through a period of cost cutting following the financial crisis. Iceland not the least. All public health care institutions in Iceland have been cutting costs. Hospitals have cut costs by 17 per cent – 18 per cent from 2008-2012. OECD has estimated the pressure on increased costs of health care in the future because of the rapidly aging population and the increasing prevalence of chronic diseases. The issue under study here is therefore highly relevant in today's health care environment.

This paper addresses the question how ageing affects health care expenditures. It has been estimated that the ratio of the population above 65 will increase significantly in the next 20-30 years. It is assumed in the paper that ageing increases the demand for health care per capita and the increase in demand for health care leads both directly and indirectly to changed expenditure. Technology is assumed to increase the level of spending but will affect the age groups differently. There is an extensive usage of explanatory variables in the model. Economic, social and demographic variables, institutional variables as well as technology and capaci-

* Ministry of Welfare, Iceland, anna.l.gunnarsdottir@vel.is.

ty variables. The variable proximity to death is not used because it is just pushed forward in time with increased longevity.

The question remains if quality of life will be improved with increased life expectancy or if a longer life means longer time w/chronic diseases and pressure on health care costs. The results of this study confirm the findings in earlier studies that GDP per capita is highly associated with health care expenditures. Unemployment has a small but statistically significant negative association with expenditure and female labour force participation has a positive association. Ageing appears to be associated with increasing health care expenditures but the association seems to disappear when income and institutional variables are included.

An additional and very interesting result in this study is that countries with salaried GPs, capitation payment of GPs, case based reimbursement for hospitals and the number of physicians per capita show a negative and statistically significant association with health care expenditures.

Most studies confirm that GDP per capita is the most important determinant of health care expenditures in the OECD countries and the effect of demographic and institutional variables is limited, but measurable. It is interesting to see that according to some studies, future health care expenditures will likely be more dependent on the costs of long-term care and medical technology and less dependent on age and gender. An OECD study from 2006 did project health care expenditures 2005-2050. In a healthy aging scenario, a change in demographics accounted for only a small increase in costs (0.6 percentage points) but in a cost pressure scenario, the costs increased by 4 percentage points. In the cost containment scenario, the cost increase was 2.5 percentage points. However the cost containment scenario in four Nordic countries projected that the costs would increase less, or 1.4-1.8 percentage points and 2.1 percentage points in Iceland. This shows the differences in health care in various countries. The quality of data is also a challenge for all research between countries. Different countries do perhaps use different accounting methods and not a standardized costs classification. Additionally, the question remains how the policy action in each country affects the total costs of health care through time.

Most empirical studies on the costs of health care in various countries use OECD data which means that countries outside OECD are not included. It would be interesting to get cost data from other parts of the world

and see if the results are similar. Also, when reading earlier research stating that health care costs for the oldest population are lower than costs incurred by other age groups, the question arises whether there exists a discrimination against old people when treatment is chosen and less or cheaper treatment is used for older people. An interesting research would be to investigate treatment patterns in different age groups. Empirical studies use historical data and it seems important to have the most current data included in the model. 2003 data is the most current in this study and it would be interesting to get data until perhaps 2010 and run the regression model again.

This paper presents an interesting and important study as well as giving an extensive overview of empirical research in this field, both macro and micro studies. The observations are clearly presented and discussed and considerations with respect to strengths and weaknesses of the study are well presented. The paper adds to the knowledge in this field and it is my hope that further studies will be done along the lines drawn in this paper.

Lifestyle, health and costs – what does available evidence suggest?

Kristian Bolin*

Summary

In this paper, the available evidence regarding certain health behaviours – smoking, alcohol consumption, nutritional choices, and physical activity – and the associated health risks and consequences for healthcare and productivity is summarized. Obesity is treated in a similar way. Moreover, the evidence for effectiveness and cost-effectiveness of primary and secondary interventions in these areas is summarized. As regards smoking and obesity, there is substantial knowledge concerning health risks, healthcare and productivity consequences. The effectiveness and cost-effectiveness of primary and secondary smoking prevention are well-known. In contrast, the current evidence concerning health risks associated with particular nutrients is in some cases ambiguous (saturated fat, certain carbohydrates, alcohol). In general, tax and information policies are effective, and may be cost-effective, in influencing health-related behaviour.

Keywords: health-related behaviour; health risk; prevention.

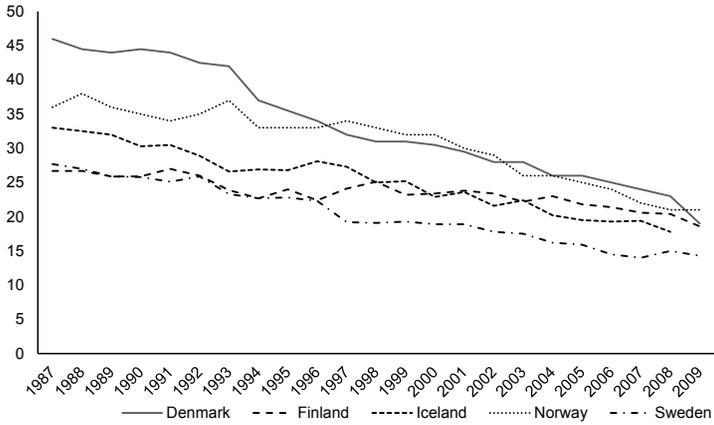
JEL classification numbers: I12; I18; I38.

* Department of Economics, Lund University, Kristian.Bolin@nek.lu.se.

Several human behaviours are subject to individual choices that entail large future health consequences. These consequences may or may not have been taken into account when, for instance, deciding to smoke, how much alcohol to drink or what and how much to eat. To the extent that all consequences of a particular behaviour have not been taken into account, there is – in principle – scope for welfare improvements. However, the current knowledge concerning causes and effects in the area of individual health-related behaviour is in many cases inadequate. Obviously, imperfect knowledge about the relationship between current behaviour and future health outcomes impairs the capacities to make efficient decisions. Likewise, an imperfect understanding of the effects of available public policy measures obstructs the construction of efficient health policies. The overall objective of this study is to provide a summary of available knowledge concerning (1) current behaviours and future health and cost outcomes and (2) the effectiveness of public policies for health. Individual choices are central for both health outcomes and responses to various policy initiatives. Therefore, I will also provide a non-technical description of the dominating human-capital framework for analysing health-related behavioural choices.

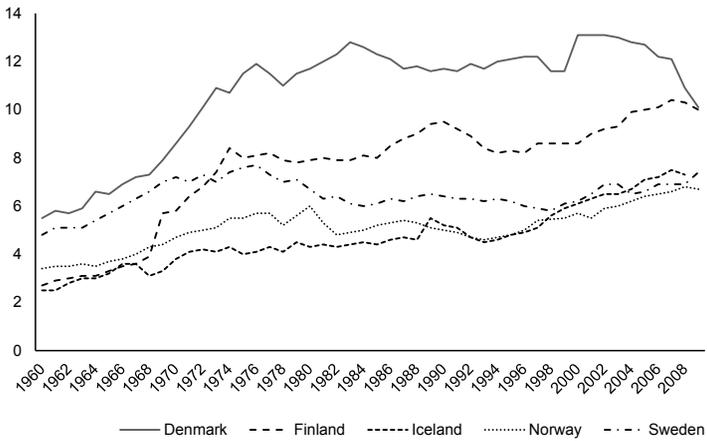
Avoidable health conditions are usually considered as mainly those caused by smoking, (excessive) alcohol consumption, poor nutritional choices, disadvantageous physical-activity levels, and illegal drug use (the problem of causation is left aside for now – a discussion of this issue can be found in the concluding section). These behaviours are related to several major diseases, for instance, circulatory and respiratory diseases, and different forms of cancer and, according to calculations published by the WHO (WHO, 2009, Table 1, high-income countries), they are responsible for more than five million annual deaths in high-income countries worldwide. More specifically, the WHO estimates suggest that 68 per cent of the deaths and 44 per cent of the disability-adjusted life years are caused by a modifiable health risk (high-income countries). The last half century has brought large population-wide changes in health-related behaviour. Some of these changes – for instance, the decrease in smoking prevalence in several countries – can be expected to have had beneficial health effects. Other changes, for instance, increased alcohol consumption in some countries and dietary changes, are likely to bring adverse health effects.

Figure 1. Smoking prevalence

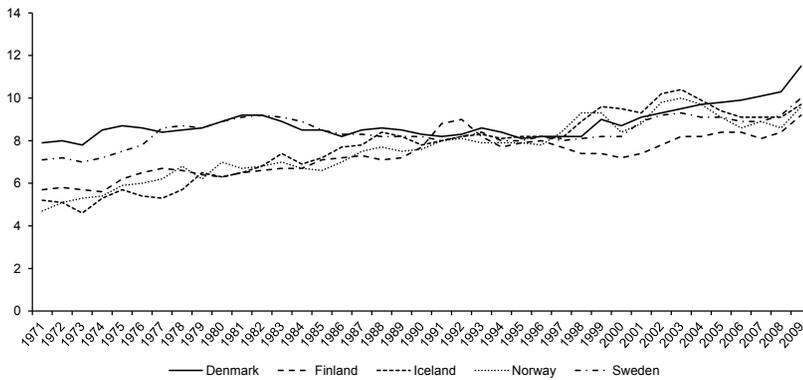


Source: www.oecd.org/health/healthdata.

Figure 2. Annual alcohol consumption, liters/cap



Source: www.oecd.org/health/healthdata.

Figure 3. Healthcare expenditures as % of GDP 1971 - 2009

Source: www.oecd.org/health/healthdata.

The Nordic countries have gone through more or less the same health-behaviour developments as other comparable countries (OECD countries). Figures 1-2 illustrate the changes in smoking and alcohol consumption and Figure 3 illustrates the development of healthcare expenditures, as share of GDP, for the time period 1971-2009. More extensive comparisons between the Nordic countries can be found in the OECD health database.¹

- Smoking
- Alcohol consumption
- Healthcare expenditures

The Nordic countries distinguish themselves as regards the prevalence of smoking, which has decreased considerably and is now lower than in most other OECD countries. Sweden is at the lower end of the range (14.3 per cent were daily smokers in 2009), while Norway is at the higher end (21 per cent were daily smokers in 2009). The most dramatic decrease in smoking prevalence is demonstrated by Denmark: in 25 years, the share of Danes that are daily smokers has decreased from about 50 per cent to less than 20 per cent. Yet, lung cancer has increased in importance as a cause of death in all countries except Finland (OECD health data).

¹ www.oecd.org/health/healthdata.

Recent evidence suggests that the benefits achieved by lower smoking-attributable healthcare costs are realized further into the future than what has previously been expected (Bolin et al., 2011).

Parallel to this development, alcohol consumption has increased in all Nordic countries. However, the countries show different developments both as regards consumption and the number of deaths due to liver cirrhosis. In Denmark and Finland, the number of deaths caused by liver cirrhosis increased by 50 and 400 (sic.) per cent, respectively, between 1960 and 2009, while the corresponding number in the other Nordic countries was about the same in 2009 as in 1960 (OECD health data).

Eating and physical exercise habits have a direct influence on health and an indirect effect through their effect on obesity. Intra-county comparable statistics as regards the development of dietary and physical exercise habits do not – to the best of my knowledge – exist (crude single-point measures are reported below). Even though the OECD health data is incomplete as regards obesity, the shares of the populations being overweight or obese can be inferred – overweight: body mass index (BMI) ≥ 25 and < 30 ; obesity: BMI ≥ 30 . The prevalence of obesity has increased, from about 6-7 per cent at the beginning of the 1970's to 10-15 per cent 35-40 years later (OECD health data).

Parallel to these developments, healthcare expenditures – as a share of GDP – rose by about 50 per cent in the Nordic countries during the period 1970-2010.

The remainder of the paper is organized as follows: the next section, Section 1, gives a brief overview of the dominating theoretical framework for health-related behaviour. The main section (Section 2) provides an account of the available knowledge concerning health effects, and the corresponding impacts on healthcare costs, of the following health-related behaviours/conditions: smoking, alcohol consumption, obesity, nutritional choices and physical activity. In Section 3, I go through the available evidence regarding the effectiveness and cost effectiveness of public policies constructed in order to modify health behaviours (primary and secondary prevention) and to reduce weight. The paper ends by a summary of the findings and an identification of important issues and challenges for future research.

The objective of this study was to provide an up-to-date summary account of (1) the relationship between the most influential health behav-

iours and their corresponding health outcomes, and (2) the scope for influencing individual health behaviours through public policy initiatives. Performing systematic reviews of the evidence published in the different research areas spanned is considered to be beyond the scope of this study. Instead, I searched for peer-reviewed articles using the Summon search engine and Lund University access to journals. Search phrases combining behaviour, for instance, smoking, with an additional topic, for instance, health risk or prevention, were used. The articles included were then chosen according to the impact factor of the publishing journal. In addition, the Cochrane Collaboration was searched for systematic reviews.

1. The demand-for-health framework

From an economics perspective, most human behaviours are manifestations of individual preferences and possibilities mediated through a decision-making process with the overall objective of making the most out of life. The last half century has brought a formidable development in the application of economics tools to more or less all human choices of significance. Although this can be traced back more than half a century, Gary Becker is maybe the most well-known contemporary economist who has applied economics to a wide variety of human behaviours, for instance, the economics of crime, and the economics of marriage, that had not previously been subject to the economist's tool kit. In 1964, Becker published his work *Human Capital*. Since then, human capital as an analytical concept has gained importance both in academic research and in more applied work and in the public debate. At least since the publication of *Human Capital*, the term human capital has been used as representing the complete set of an individual's capabilities. In practice, education – both formal schooling and on-the job training – was the dominating human-capital component in most studies during the 1960's and the 1970's.

Human-capital research put a particular focus on health in 1972 when Michael Grossman published his extension of Becker's theory of human capital (Grossman, 1972, 2000). The fundamental notion in his extension is that health has properties distinguishing it from all other components of the human-capital stock – health does not only affect the efficiency of the individual in different activities, it also influences the amount of time that

the individual can divide freely between different productive purposes. Further, Grossman's treatment of health means that the individual can influence his or her own health by making investments in the health stock. This did not only constitute a novel approach to health itself, but also to the demand and utilization of healthcare goods and services. In Grossman's model of health and investments in health, it followed naturally that healthcare is utilized not for its own sake, but because of its effect on health. In other words, the demand for healthcare is derived from the more fundamental demand for good health.

The demand-for-health model is the foundation of much current research on health-related behaviours. It follows the neoclassical economics tradition in that its point of departure is the individual making choices in accordance with his or her preferences, and subject to restrictions as regards individual capabilities and resources. It should be emphasized that this view of individual behaviour does not preclude that many actions may have both positive and negative effects on health, as long as the individual's total wellbeing increases. Obvious examples are smoking, which clearly involves serious future health hazards, but increases the current wellbeing of the smoker, and alcohol consumption, which may be harmful both in the short and the long run but, depending in the level of consumption, may provide both current welfare and long-term beneficial health effects. Several other behaviours are both subject to individual choice and related to health, for instance, physical exercise, eating, sleeping etc. Thus, what is usually labelled "lifestyle" (or health-related lifestyle) is, as far as its health effects are concerned, made up of several different behaviours which may or may not influence health in the same direction.

The essentials of the demand-for-health model are (a non-technical presentation of the demand-for-health model can be found in Bolin, 2011):

- the individual uses his or her resources in order to obtain the highest possible welfare over the course of life;
- health is regarded as a capital good which improves welfare but requires costly investments to prevent too low levels;
- the individual "produces" his or her own investments in health, using healthcare, market goods and own time;

- the amount of investments produced at each stage of the lifecycle depends on the individual's (a) relative valuation of health and other commodities, (b) degree of impatience regarding intertemporal choices, (c) capability of transforming knowledge about the health consequences of specific behaviours and tangible resources into health investments; and
- it is assumed that the individual has perfect knowledge about the future health consequences of her current behaviour.

The great importance, and large policy relevance, of human capital and individual decisions that determined human capital has been emphasized several times in the literature. This is further accentuated by the ongoing demographic transition that makes the future work ability of those below retirement age relatively more important. Thus, knowledge about (1) individual characteristics and health-related decisions, (2) the mechanisms relating individual decisions and health outcomes, and (3) the possibility of influencing individual health-related decisions through various policy measures, is immensely important for the current and future welfare of societies. The demand-for-health model provides guidance for research in all these areas. There are, however, several aspects of health-related behaviour that need further theoretical and empirical attention.

A notable criticism of the original formulation of the demand-for-health model concerns the lack of certain real-life aspects that are important for health-related decisions. The theory has been extended in order to include some of these aspects: uncertainty (e.g., Liljas 1998, 2000), the family as a producer of health (Bolin et al. 2001, 2002a), the importance of labour market conditions for health (Bolin et al., 2002b) and the importance of healthy and unhealthy consumption (Forster, 2001). A more technical criticism of the demand-for-health model that has been mentioned in the literature concerns the fact that health is treated as endogenous to the degree that the individual in effect decides about his or her time of death. Empirical estimates of the demand-for-health model have not unambiguously agreed with the theoretical predictions derived from the model. This may be explained by most of the studies having used cross-sectional data (Bolin, 2011), which cannot capture the dynamic features of the theoretical demand-for-health model (for a survey of

empirical estimates based on the demand-for-health model, see Grossman, 2000).

In addition to the specific attributes of health-related behaviours captured within the demand-for-health framework, various other real-world complications to the decision-making process behind health-related behaviours have been highlighted in the literature: the willingness to trade-off current for future welfare may be time inconsistent (hyperbolic discounting; see, for instance, Laibson, 1997); there may be peer-effects determining, for instance, choices to consume addictive goods (see, Becker and Murphy, 1988) and certain behaviours may have a U-shaped (Bolin and Lindgren, 2012) association with health etc. In principle, effectiveness (and cost-effectiveness) requires that these complications are taken into account when designing public health policies.

1.1 Foundations for public health policies

Health-related behaviours may be inefficient – the true costs are not balanced against the true benefits – for at least two reasons: poor decision making with respect to the utilization of available knowledge and/or external effects. External effects may arise whenever, as in most societies, the individual does not face the full consequences of his or her health-related behaviour. Poor decision making can be further classified according to whether perfect information is (in principle) available and whether available knowledge is fully utilized by the individual. Thus, in as much as the adverse health consequences that result from smoking, alcohol misuse, poor diet and too little physical exercise are not taken into account when individuals decide about their actions, society as a whole may be better off if those behaviours are modified in order to balance total costs and benefits. External costs are likely to arise for all behaviours considered in this study (Zohrabian and Philipson, 2010; Cawley and Ruhm, 2011). Thus, public policy measures with the objective of achieving behavioural changes seem to be well-founded from a welfare perspective – with perfect knowledge concerning external costs, an efficient solution to the externality problem would be to charge, for instance, each smoker the external cost that he or she does not pay for. However, in practice, the external effects are not known with a large enough precision to make this possible. Instead, public efforts that combine rather arbitrary

payments from smokers – taxes – are combined with interventions with the objective of reducing smoking, or preventing young people from starting to smoke. All these efforts are costly and, hence, the resources used have to be balanced against the benefits that accrue, and if the net benefit associated with a particular intervention is positive, that intervention should be implemented. Such cost-benefit analyses are often infeasible due to lack of information regarding individual demand. In practice, the issue of whether to implement a specific intervention is mostly settled by comparing, on the one hand, induced costs and benefits, with the willingness to pay for, for instance, a quality-adjusted life year. This is referred to as a cost-effectiveness analysis, and whether or not an intervention is cost-effective depends on the willingness to pay for achieving certain effects. For instance, if the willingness to pay for an additional QALY is EUR 50 000, interventions which produce an additional QALY at a net-cost below EUR 50 000 are cost-effective.

2. What is known about the influence of individual behaviour on health outcomes and costs?

It is well-known that individual behaviours do influence future health outcomes. For instance, smoking, alcohol abuse, drug use and poor diet may lead to bad health and an increased risk of premature death. However, available epidemiological knowledge concerning the relationship between, on the one hand, different health-related behaviours and future health outcomes and healthcare costs, on the other, is far from complete. *First*, several published studies suffer from a poorly established causality between the studied behaviour and health and healthcare utilization, respectively (Cawley and Ruhm, 2011). Whenever causality has not been established, uncertainty arises as to whether certain health policies will have the intended effects. Strictly speaking, this means that the policy effort taken in order to enhance public health by altering a specific (assumed-to-be) health-adverse behaviour may be efficient in changing that behaviour but, if causality runs in some other direction than that assumed by the policy maker, it may nevertheless not have the desired effects on health. Obviously, this constitutes a serious problem for policy makers in the health area. *Second*, while the health effects of, for instance, smoking

can be said to be firmly established, the available knowledge concerning the relationships between alcohol consumption and nutritional choices is inadequate or even confusing. *Third*, even though the causality between a particular behaviour and health may be considered as established, the magnitude of the relationship may still be largely unknown. In particular, for specific sub-groups not represented in the studies that demonstrate causality, the evidence for causality may be poor as regards both quality (the presence of a causal effect) and quantity (the size of the effect).

The method of randomly assigning individuals into two groups – one to which treatment is provided and another to which no treatment is provided – is the most convincing method for establishing cause and effect. New medical treatments, for instance, new pharmaceutical-based treatments, are frequently subject to clinical trials designed as randomized controlled trials. However, this method is not feasible in many practical situations involving behaviours that are chosen on an individual level and a day-to-day basis. This means that researchers often have to rely on natural experiments in order to study the relationship between cause and effect. A natural experiment is a method that explores exogenously imposed differences between groups in order to study the relationship between, for instance, health-related behaviours and health outcomes. Examples of natural experiments are (1) when two comparable populations are subject to different vaccination policies or (2) when two comparable populations are subject to different health information policies. Neither the randomized experiment nor the natural experiment is a perfect method for inferring the causality between particular actions and outcomes, though. The validity of the results from a randomized experiment study may be questioned since the population studied may poorly represent the population at large, that is, the results may lack in external validity. Likewise, the natural experiment may provide information from which causality is hard to determine. Cawley and Ruhm (2011) summarize the problems with establishing causality in different research designs.

As mentioned above, empirical evidence suggesting that the relationships between health behaviours and health outcomes may be less straightforward than previously assumed has emerged during the last 15-20 years. In particular, recent empirical findings suggest that certain alcohol (wine) consumption patterns may entail health benefits (Gaziano et al., 1993; Thun et al., 1997) and that the optimal mix of nutrients may

differ from the established beliefs. The evidence is mixed in both cases. Naturally, lifestyle choices, as well as policy efforts with the objective of enhancing public health, are both obstructed when the relationships between behaviour and outcomes are unknown or when decisions are made using erroneous information. This means that individual decisions regarding health-related behaviours may be at a certain distance from what they should have been in a world with perfect knowledge. Thus, observed health-related behaviours are likely to be poor indicators of underlying preferences in those cases when decisions have been made using erroneous information. Moreover, improved information about health risks will not always induce the intended modifications in individual behaviour – the direction of the behavioural modification will depend on whether health risks were over- or underestimated prior to the new information.

It is not the occurrence of illness per se that distinguishes population groups according to health behaviour – it is the rate at which illness and death occur. Thus, not all cases of ill health can be attributed to specific health behaviours, which has to be taken into account in analyses of future health outcomes and current behaviour. For instance, some people get lung cancer or COPD (chronic obstructive pulmonary disease) even though they have not smoked a single cigarette in their lives. There are two main empirical methods for taking this into account and, hence, for inferring the consequences for future health and healthcare costs and lost productivity of a specific current health-related behaviour. *First*, the epidemiological approach utilises epidemiological evidence in order to calculate attributable fractions. The attributable-fraction method (for a non-technical presentation of the joint effect of multiple risk factors see, for instance, WHO, 2009) provides a way of deriving estimates of the proportions of specific diseases that can be attributed to the studied behaviour, using the health risks and the population prevalence that correspond to the studied behaviour. The method produces a percentage for each included health condition, which indicates the proportion of all cases of the condition that can be attributed to, for instance, smoking. Thus, these proportions also reflect the amount of healthcare utilisation that can be attributed to this behaviour. Below, the relative risks associated with specific behaviours and health conditions are reported. *Second*, various statistical methods are available that make use of data that links health out-

comes and health-related behaviour in order to infer the “pure” causal effect of the studied behaviour.

This study will focus on the health effects of smoking, alcohol consumption, nutritional choices and physical activity. The literature regarding smoking and its consequences is more extensive than what is the case for the other behaviours that I consider. Moreover, I will focus on the health risks associated with *being* obese. Diet and activity choices largely constitute the risk-factors for obesity.

2.1 Smoking

In spite of various policy efforts around the world, the prevalence of smoking remains high in most countries: in Western Europe, between 15-60 per cent are daily smokers; the corresponding figure for the United States is 14-20 per cent; in Eastern Europe the prevalence of smoking is even higher, for instance, 65 per cent of Russian men are daily smokers (World Health Organization, 2009). The Nordic countries show somewhat lower prevalence rates of smoking than the rest of Western Europe – between 15 and 20 per cent are daily smokers. Estimates suggest that smoking caused about 5 million premature deaths per year worldwide at the beginning of the twenty-first century, and within the next ten years, the estimated annual number of premature deaths due to smoking would be approximately 9 million (Shafey et al., 2003). At the same time, smoking-cessation therapy has been argued to be underutilised as a means of lowering the smoking prevalence (Hughes, 2010). Although the utilisation of smoking-cessation treatment is certainly determined by a wide range of circumstances, inadequate reimbursement and less than perfect information about the effectiveness of smoking-cessation treatments among smokers have been suggested in the literature as important factors that are within reach of additional policy efforts (Abrams et al., 2010; Hughes, 2010; Levy et al., 2010).

Effects of smoking on health – the epidemiological evidence

Health risks associated with tobacco smoking include several major diseases, for instance, different types of cancers, cardiovascular diseases, and respiratory diseases. Not only do smokers face elevated risks of being struck by one or more of these diseases, the risks are substantial as com-

pared to the corresponding risks faced by those who have not ever smoked (Thun et al., 2000) and, hence, smoking is likely to be the single most important preventable health risk in the world. For instance, smokers face more than 20 times as high a risk of dying due to lung cancer than those who have never smoked. The corresponding relative risks for COPD and stroke and ischemic heart disease are about 10 and 2. Adverse health effects may even occur for non-smokers; passive exposure increases the risk of lung cancer, heart disease, and respiratory illness (Fagerström, 2002). The detrimental health effects of smoking have been known at least since the 1950's (Bolin and Lindgren, 2007; Royal College of Physicians, 1962). The epidemiological evidence has been refined over time, though. It is now possible to derive firm estimates of how large a share of morbidities and mortality that can be attributed to smoking. Estimates reported by the WHO suggest that in high-income countries, about 18 per cent of the deaths are due to smoking. For the Nordic countries, this figure is likely to be somewhat lower: a recent study performed for Sweden indicates that about 10 per cent of the deaths in the age group 35-85 can be attributed to smoking.

Using available epidemiological evidence together with information about smoking prevalence, it is possible to calculate attributable proportions for a number of diseases. For Sweden, proportions of diseases that can be attributed to smoking have been reported by Bolin et al. (2011). The proportions vary between 70 and 80 per cent for lung cancer and COPD, and between 20 and 50 per cent for different types of heart disease. Considering the development of smoking-prevalence rates in Denmark, Finland, Iceland and Norway, respectively, the share of the incident cases of each considered disease is even higher in all those countries.

Cost of smoking

Several studies have been published on the costs imposed on society by smoking that apply the attributable-fraction method. For Sweden, three studies have been published since 1985 (Hjalte et al., 1985; Bolin and Lindgren, 2007; Bolin et al., 2011). The recent study performed for Sweden reports the somewhat surprising result that the annual number of deaths that can be attributed to smoking is not decreasing, in spite of the decrease in smoking prevalence. This suggests that the adverse health effects of smoking among former smokers do not decrease as quickly as

expected. Calculations for the year 2007 show that the costs attributable to smoking amounted to about 0.4 per cent of GDP in 2007 (sickness-absenteeism from work not included). This cost included 1 227 lost life years, per 100 000 inhabitants.

Studies that utilize individual data to directly estimate the causal effects of smoking – without taking the detour around attributable fractions – have estimated the effect of smoking on sickness absenteeism, healthcare and early retirement. The effect of smoking on productivity related to short-term absenteeism from work is substantial. Using Swedish data, Lundborg (2007) showed that smokers have about 8 additional sick days per year as compared to non-smokers. Lundborg's results corroborated the earlier results obtained by Roberts and Lindgren (2001). They also estimated – once more using Swedish data – the impact of smoking on healthcare utilization and the incidence of early retirement. The relative risk of early retirement for smokers compared to never smokers was estimated at about 6.3.

Rasmussen et al. (2004) estimated the total lifetime costs of smoking, using the simulation approach and a Danish setting. They found that smoking imposes costs on society, also when a smoking-induced reduction in life expectancy is taken into account and, hence, the notion that a shorter life expectancy balances higher healthcare expenditures among smokers is refuted.

2.2 Alcohol consumption

In Western Europe, about 80 per cent of the population consume alcohol (WHO, 2009; Table A2, Europe, high-income countries). According to the same source, the prevalence of individuals consuming more than 40 grammes of alcohol per day is about 25 per cent. The trend in the Nordic countries, during the last 25-30 years, has been towards higher alcohol consumption. The exception is Denmark where there is a tendency to decreasing consumption. Up-to-date estimates suggest that Danes and Finns consume about 10 litres per capita per year, while people in Iceland, Norway and Sweden consume about 7 litres per capita per year. Excessive alcohol consumption causes 25 000 deaths annually in European high-income countries (WHO, 2009; Table A3). At the same time, there is evidence suggesting that moderate alcohol consumption is advan-

tageous for health (see, for instance, Thun et al., 1997). This finding has been challenged by, for instance, Liang and Chikritzhs (2010) who argue that the beneficial finding of moderate alcohol consumption may be due to self selection – those suffering from certain health conditions were found to be more likely to have stopped or reduced their alcohol consumption.

Effects on health of alcohol consumption – epidemiological evidence

Health risks associated with excessive alcohol consumption include several diseases, for instance, different types of cancers and cardiovascular diseases. More specifically, those who utilise alcohol excessively (more than 4 drinks per day) face higher risks of dying due to liver cirrhosis (the relative risk, in comparison to abstainers, is 7.5) and alcohol related cancer (the relative risk = 2.8) (Thun et al., 1997). A recent study presents higher relative risks for these conditions: 9.5 and between 3.6 and 5.4, respectively (Rehm et al., 2003). The relative risks for other diseases, such as cardiovascular diseases, are much lower and, hence, in as much as they are preventable, have other main causes than alcohol consumption. There is some evidence that a moderate alcohol consumption lowers the risk of coronary heart disease (see, for instance, Mente et al., 2009).

Available epidemiological evidence concerning health risks associated with alcohol consumption makes it possible to calculate country-specific attributable proportions for relevant alcohol related diseases. As demonstrated above, the relationship between health risks and alcohol consumption is less straightforward than the corresponding relationship for smoking – smoking is never beneficial for health, while a moderate alcohol consumption may be. Thus, when computing attributable proportions for a specific population, specific consumption patterns must be taken into account, that is, information about utilisation among the alcohol consuming population. Published estimates suggest, indirectly, that the proportion of alcohol-attributable cases of liver cirrhosis in the Nordic countries is about 75 per cent of all cases (Rehm et al., 2009). The number of deaths due to liver cirrhosis in the Nordic countries reflects the development of drinking behaviours: Denmark and Finland show both considerably higher alcohol utilisation rates and liver cirrhosis deaths. In the other countries, the number of deaths due to liver cirrhosis has remained more or less constant for the last 30 years.

Costs of alcohol consumption

Jarl et al. (2008) calculated the societal cost of alcohol consumption in Sweden in 2002. Their calculations suggest that that alcohol consumption imposes a societal cost at about 0.9-1.3 per cent of GDP. They estimated that almost 30 000 potential life years were lost due to alcohol attributable deaths in 2002 in Sweden – the corresponding number of life years for smoking has been estimated at about 90 000 (see above). The figures reported by Jarl al. are somewhat lower than the estimate reported by Rehm et al. (2009) for high-income countries (average).

No studies of costs identifying the causal cost effects of alcohol consumption in the Nordic countries were identified. Cawley and Ruhm (2011) mention a few such studies, for instance Balsa et al. (2008) who examined the relationship between alcohol consumption and healthcare utilization. He found that moderate drinking decreases the likelihood of emergency visits and hospitalizations (women, not men).

2.3 Nutritional choices

Nutritional choices influence future health. Specific health conditions, such as obesity, diabetes, different types of cancers and cardiovascular diseases are usually considered as being (partly) caused by past eating and physical activity habits. Likewise, obesity is caused by inappropriate dietary habits in combination with insufficient physical activity.

It should be stated already at the outset of this section that the epidemiological evidence, as regards health risks associated with eating habits, is not as conclusive as the corresponding evidence for smoking or excessive alcohol consumption. In other words, what constitutes a healthy diet is less clear than what constitutes a healthy attitude towards smoking and drinking. In contrast, the reported health risks associated with obesity and diabetes are more or less conclusive, at least as regards the qualitative content – different sizes of the perceived risks are reported in the literature. In this section, I will report on published health risks associated with certain dietary components and with being obese. Before going through that evidence, however, I will provide an account of the various risk factors behind obesity.

Although the causes of obesity seem rather obvious – an unfavourable, or individually inappropriate, balance between energy intake and

physical activity – recent research suggests that also the qualitative content of the energy intake is important (Kitahara, 2010). More specifically, the consumption of high glycemic index carbohydrates has been suggested as increasing the risk for obesity (Roberts, 2000).

Fulponi (2009) demonstrate that the supply of dietary calories per capita has increased significantly in industrialized countries since the second half of the 1900's (Sweden is the only Nordic country for which estimates are reported). More qualitative measures of observed nutritional choices and nutrition-related related risk factors, for instance, high blood pressure and high cholesterol, pertaining to high-income European countries are presented in the WHO Global Health risks report (WHO, 2009).

Health risks associated with too small amounts of dietary fruit and vegetables

It has been generally considered, and stated in public dietary guidelines, that eating too small amounts of fruit and vegetables, saturated fat, and too much red meat, is unhealthy. The epidemiological evidence that a diet that contains fruit and vegetables reduces the risk for cardiovascular diseases is firm (Estaquio et al., 2008; Fulponi, 2009). The evidence that consumption of fruit and vegetables conveys protection against cancer is less solid, though. A review of the evidence that fruit and vegetables protect against cancer was performed by a panel of researchers and published by the World Cancer Research Fund (2009). The review found that fruit and non-starchy vegetable consumption yields protection against some cancers, such as that of the mouth, larynx, stomach, oesophagus, and pharynx. However, some studies did not find any evidence for a protective effect following fruit and vegetables consumption (see, for instance, Takachi et al., 2008).

The relative risks for cardiovascular diseases and some cancers, associated with increasing the consumption of fruit and vegetables by one standard serving per day, (80 grammes) were reported by Lock et al. (2004). The relative risks are between 0.90 and 0.99 for most health conditions. Thus, the effects are small but significant. Somewhat lower risks were reported by Mente et al. (2009), who performed a systematic review of the evidence supporting a causal link between diet and coronary heart disease. They concluded that dietary fruit and vegetables reduce the risk for coronary heart disease.

Health risks associated with saturated fat

The epidemiological evidence that consuming saturated fat increases the risk of cardiovascular diseases and cancers has recently been questioned and new evidence pointing in the opposite direction has emerged. In a recent Cochrane review (Hooper et al., 2011), it was concluded that there is evidence that a reduction in the consumption of saturated fat reduces the risk for cardiovascular diseases. The relationship is weak, though. However, several studies report no effect of reducing saturated fat in the diet (Siri-Tarino, 2010a), some even reported an increased risk of doing so (Siri-Tarino, 2010b). Mozaffarian (2011) argues that there is convincing evidence that saturated fat does not increase the risks of coronary heart disease and cancers, respectively. The recent meta-analysis of prospective cohort studies by Siri-Tarion et al. (2010a) concluded that there is no significant evidence that dietary saturated fat is associated with cardiovascular disease, which corroborated the findings reported by Mente et al. (2009).

Health risks associated with high glycemix-index carbohydrates

Carbohydrates with a high glycemic index have been shown to increase the risks for cardiovascular diseases (for reviews, see Siri-Tarino et al., 2010b; and, for coronary heart disease, Mente et al., 2009). Levitan et al. (2007), who performed a study on Swedish males, did not find any association between dietary glycemic index and dietary glycemic load, respectively, and ischemic cardiovascular disease. However, they found that glycemic load increased the risk for a hemorrhagic stroke. In a later study (Levitan et al., 2009), no association was found between cardiovascular and all-cause mortality and dietary glycemic index and load, respectively, among men with an established cardiovascular disease. Thomas et al. (2007) reviewed the evidence that diets containing low glycemix-index carbohydrate are more efficient in reducing weight among obese individuals than diets containing carbohydrates with a higher index. They included six randomized controlled studies and concluded that low glycemix-index diets are more effective in reducing weight. Kitahara (2010) reviewed and discussed the evidence for an association between low-glycemic load diets and diabetes, cardiovascular disease and some cancers. She concluded that in spite of the methodological limitation of the

available studies, there is a growing stock of evidence that the low-glycemic load diets decrease the risks associated with these diseases.

Health risks associated obesity

The most recent published relative risks associated with being obese suggest that obesity significantly increases the risk for several cancers, for instance, cancer of the esophagus, the colon, and the liver (Renehen et al., 2008). The corresponding relative risks are 1.5, 1.2 and 1.2. Even though these risks are low as compared to the relative risks for various diseases faced by smokers, several cases attributable to obesity will still occur since the prevalence of obesity is relatively high. Gelber et al. (2008) report the relative risks for cardiovascular disease associated with being obese. The risk fell in the interval 1.4-2.6, depending on the specifics of the statistical method utilized. Similar figures were reported by Ni et al. (2004).

Costs of obesity

A recent review (Withrow and Alter, 2011) of the worldwide economic burden of obesity identified and included only 1 study regarding a Nordic country – Sweden. It was concluded that obesity accounted for between 0.7 and 2.8 per cent of the total healthcare expenditures in any given country. Finkelstein et al. (2010) calculated the expected years of life lost due to obesity, using US data. They found that the adverse influence of obesity on length of life depends on smoking status and age when diagnosed as obese. Young severely obese (BMI >40) smokers are expected to lose between 7 and 12 life years, while elderly severely obese smokers face a reduction in life expectancy of between 4 and 6 years. Further, they found no association between overweight and mild obesity ($30 < \text{BMI} < 35$) and reduced life expectancy. Trogon et al. (2008) reviewed the evidence for an association between obesity and cost related to lost productivity. They found evidence that obese workers are more absent due to illness than their non-obese peers. They also found that the costs associated with premature mortality vary greatly between countries.

The number of potential life years lost due to obesity and overweight in Germany was recently calculated (Konnopka et al., 2011). About 428 000 potential life years were lost in Germany due to obesity and overweight in 2002. The prevalence of obesity and overweight is about

the same in Sweden and Germany and, hence, more than 40 000 potential life years are likely to be lost due to obesity and overweight in Sweden. This estimate places obesity beyond both smoking and alcohol abuse as a cause of loss of life, a conclusion that corroborates the findings by Sturm (2002).

2.4 Physical activity

In studies of physical activity, individuals are often characterized as belonging to one of a finite number of groups, according to the level of activity. The epidemiological evidence that physical activity entails substantial beneficial health effects is firm. However, the exact amount and composition of physical activity required in order to achieve potential health enhancements are less clear. WHO (2009) provides consistent estimates of shares of different populations according to physical activity. For high-income European countries, almost 70 per cent of the population are inactive or insufficiently active. This estimate is consistent with estimates derived from the Swedish Survey of Living Conditions (Bolin and Lindgren, 2006).

Effects on health of inadequate physical activity – epidemiological evidence

The relative risks for cancer and cardiovascular disease, associated with being inactive versus being active, were reported by Garrett et al. (2004): the relative risks for breast- and colon cancer are 1.5 and 2, and the relative risks for ischemic heart disease and stroke are both 2. Garrett et al. report population attributable risk proportions for cancer and cardiovascular disease, respectively, at about 30 per cent. Using physical-activity prevalence measures for Sweden gives somewhat lower attributable proportions.

Costs of inadequate physical activity

Bolin and Lindgren calculated healthcare costs and lost productivity in Sweden in 2002, due to physical inactivity or insufficient activity. They concluded that 0.3 per cent of the total healthcare costs and about 27 000 potential life years lost were due to inadequate physical activity. The

estimated loss of potential life years is on par with the corresponding estimate for alcohol consumption.

3. What is known about the effectiveness of public health policies?

The ultimate goal of public health policies is to improve health. In principle, this can be achieved either by influencing behaviours that affect health risks – prevention – or by supplying subsidized healthcare in the event of disease. I shall focus on the available means of influencing individual health behaviours and thereby preventing adverse health outcomes.

Both primary prevention – preventing the adverse health behaviour from ever occurring – and secondary prevention – altering the behaviour once it has occurred – methods have been used in all areas discussed above. Below, I report on the effectiveness of both primary and secondary prevention methods, focusing on published systematic reviews of the available evidence. For tobacco, alcohol and obesity, respectively, both primary and secondary prevention methods have been evaluated. However, the distinction between primary and secondary prevention is not equally useful for physical activity and nutritional choices. In these two cases, I have considered all prevention as secondary prevention, since most – if not all – interventions focus on altering unhealthy behaviour rather than preserving healthy behaviour. An account of the published evidence as regards the cost-effectiveness of public health policies concludes the section.

Cawley and Ruhm (2011) provide a summary of the evidence on the effectiveness of different policy measures in influencing health behaviours. Tax and information policies are efficient in influencing health-related behaviours. However, tax policies that target eating habits may be subject to difficulties since several foods cannot be unambiguously divided according to health effects.

3.1 Primary prevention

Prevention in order to restrain young people from starting smoking may be effective, but the evidence is weak (Lantz et al., 2000; Thomas and

Perera, 2006; Brinn et al., 2010; Chaloupka et al., 2011). Lantz et al. (2000) concluded that youth smoking prevention efforts have shown a mixed effectiveness. However, they outlined several promising specific interventions that warrant an additional study, for instance, as regards media campaigns, community-based interventions and price initiatives. The two Cochrane reviews – Thomas and Perera (2006) and Brinn et al. (2010) – report on the accumulated evidence regarding the effectiveness of school-based programmes and media campaigns for preventing young people from starting smoking. Thomas and Perera (2006) identified 23 randomized control studies of high validity. 9 studies demonstrated that social influence is effective in preventing young people from becoming smokers. However, they also concluded that there is little evidence that information alone will prevent non-smokers from starting smoking. The most recently performed review focused on media campaigns and their effectiveness in preventing young people from starting smoking (Brinn et al., 2010). They identified and included 7 studies from which they concluded that there is some evidence that such campaigns are effective, but the evidence is weak and methodologically flawed. Finally, Hajek et al. (2009) reviewed the effectiveness evidence for relapse prevention methods, targeting recent quitters. They identified and included 36 studies from which they concluded that interventions based on providing individual skills needed to avoid relapse have no effect. However, extended pharmacological treatment – varenicline or nicotine replacement therapy beyond the smoking-cessation treatment – was found to significantly reduce the risk of relapse.

Foxcroft and Tsertsvadze (2011) reviewed the evidence for the effectiveness of family-based programmes designed to prevent alcohol misuse among school children (<18 years of age). 12 trials were included in the review. They concluded that there is some evidence that family-based alcohol prevention programmes targeting children are effective. However, the results from the trials were poorly reported, which reduced the perceived validity of the results. Anderson et al. (2009) reviewed the evidence for the effectiveness (and cost-effectiveness) of policies designed to reduce harm caused by alcohol abuse. They conclude that policies targeting prices, availability and banning alcohol advertisement are effective.

Interventions that target diet and physical activity habits have been evaluated as tools for obesity prevention. The evidence for the effective-

ness of interventions designed with the objective of preventing the development of obesity in children was recently reviewed by Waters et al. (2011). They included 55 studies, the majority of which targeted children in the age span 6-12. A meta-analysis was performed using data from 37 of the 55 studies. The result suggests that obesity prevention programmes designed to prevent obesity among children, aged 6-12, are effective. Although the authors found the evidence to be strong, they argued that the results should be interpreted with caution, due to methodological problems in the underlying trials. The included studies employed a broad variety of different programme components and the reported results provide little guidance as to the relative contribution of each individual component. Waters et al. argue that their analysis identifies specific components to be the most potent ones. They provide a list of 6 specific components, all based on school and family support.

3.2 Secondary prevention

The evidence for the effectiveness of various secondary prevention programmes is considerably more extensive than the evidence for the effectiveness of primary prevention programmes. In particular, a broad variety of smoking-cessation interventions have been evaluated.

Several systematic reviews of the effectiveness of smoking-cessation interventions have been published. Wakefield and Chaloupka (2000) and Grimshaw and Stanton (2006) reviewed the evidence for smoking-cessation interventions targeting young smokers. Wakefield and Chaloupka found that combinations of public information campaigns and price policies are effective in reducing smoking among both adolescents and adults. Grimshaw and Stanton included 24 trials in their review, and found some evidence that complex interventions, involving several different components, may be effective. The available evidence, however, is not sufficient for the design and implementation of particular programmes. There is no evidence that pharmaceutical-based interventions are effective for adolescent smokers. Pharmaceutical-based interventions mainly utilize three different drugs (bupropion, nicotine replacement, and varenicline). The evidence for the effectiveness of interventions based on these pharmaceuticals was reviewed by Stead et al. (2008) and Cahill et al. (2011).

The Stead et al. review included 132 trials that compared NRT (nicotine replacement therapy) either with a placebo or a non-nicotine alternative. They found firm evidence that smoking-cessation therapy using nicotine replacement increases the chance that a quit attempt is successful by 50-70 per cent. Further, it was concluded that these effects are independent of supporting counselling. The Cahill et al. review included 11 trials that compared varenicline with bupropion, NRT, and placebo. They performed a meta-analysis and found that varenicline instead of no pharmaceutical support (placebo) more than doubled the chance that a quit attempt is successful. Compared to bupropion and NRT, the benefits of varenicline were less pronounced: the chances of success increased by 52 and 13 per cent, respectively.

Internet-based smoking-cessation interventions were evaluated by Cijljk et al. (2010). They included 20 studies and found evidence that internet-based programmes in combination with other interventions, for instance, NRT, may be more effective than either intervention alone. Moreover, Cijljk et al. argued that the benefits of internet-based interventions are their potential for providing individually tailored information, and that they may be relatively attractive to young people.

As regards alcohol misuse, both pharmaceutical-based interventions and interventions based on the social context have been developed and evaluated. Rösner et al. (2010a, 2010b) evaluated interventions based on pharmaceuticals and constructed in order to curb alcohol dependence. The evidence for the effectiveness of social-context interventions was evaluated by Moreira et al. (2009). Rösner et al. (2010a) included 24 studies and performed a meta-analysis. They found that treatment with acamprosate is effective for maintaining abstinence from alcohol consumption. It was also concluded that acamprosate treatment in combination with psychosocial programmes further increased the effectiveness of the intervention. Further, it was argued that the moderate treatment effect (the risk of alcohol consumption decreased by 14 per cent) should be appreciated in the light of the lack of alternative treatments and the reoccurring nature of alcoholism. Rösner et al. (2010b) evaluated the effectiveness of alcoholism treatment using naltrexone. 50 randomized controlled studies were included and a meta-analysis was performed. They found results close to those reported by Rösner et al. (2010a).

The effectiveness of social-norm feedback in reducing alcohol misuse was evaluated by Moreira et al. (2009). They found some evidence that feedback using either internet-based or individual face-to-face feedback may be effective in reducing alcohol misuse.

Compared to interventions designed with the objective of altering tobacco and alcohol habits, interventions targeted at improving nutritional choices and physical activity have been more scarcely evaluated. Dobbins et al. (2009) reviewed and evaluated the evidence regarding the effectiveness of school-based interventions designed with the purpose of promoting physical activity among children and adolescents (aged 6-18). 26 studies were included and it was concluded that school-based programmes are effective in increasing the amount of physical activity undertaken by those who are physically active, but have no influence on the decision whether or not to be active. Dobbins et al. recommend that school-based programmes for promoting physical activity are implemented.

Brunner et al. (2007) reviewed the evidence for the effectiveness of providing dietary advice in order to achieve sustained dietary changes or an improved cardiovascular risk profile among healthy adults. They identified and included 38 trials and concluded that providing dietary advice is effective in bringing about advantageous changes in diet and cardiovascular risk. They assessed the effects over a ten-month period.

Interventions designed in order to promote a more physically active lifestyle among sedentary adults were evaluated by Foster et al. (2005). They included 19 studies from which it was concluded that interventions comprised of professional advice and support may influence sedentary adults to become more physically active. The effect is moderate, however. The available effectiveness evidence regarding community-wide, multicomponent, interventions with the objective of increasing the population levels of physical activity was evaluated by Baker et al. (2011). They included 25 studies and concluded that there is no evidence in support of the effectiveness of this type of interventions in promoting physical activity. Suggestions for future studies include an improved design and measures of outcomes and larger samples of participants.

Interventions that target obesity have been evaluated by Oude et al. (2009), Curioni and André (2006) and Padwal et al. (2003). Oude (2009) evaluated lifestyle and pharmaceutical-based interventions. They identi-

fied and included 64 randomized controlled trials. Meta-analyses were performed and it was concluded that lifestyle interventions are effective for reducing overweight among both children and adolescents. Adjunctive pharmacological treatment for severely obese adolescents may be beneficial. Curoni and André evaluated interventions based on drug treatment (rimonabant). They identified and included four studies, from which it was concluded that treatment with rimonabant has small effects on weight. However, adverse effects were also reported. Padwal et al. evaluated the long-term effect (> 1 year) of treatment with anti-obesity pharmaceuticals. They included 30 studies (16 orlistat trials; 10 sibutramine trials; and four rimonabant trials), and concluded that all three anti-obesity agents studied are moderately effective in reducing weight. The validity of the included studies was limited due to high attrition rates, and it was concluded that longer and more methodically rigorous studies are required.

3.3 Cost-effectiveness

The cost-effectiveness of primary prevention health interventions has been much less studied than the cost-effectiveness of secondary prevention methods. No doubt is this due to the greater difficulty in assessing the effect of a typical intervention in the first case. Therefore, in this section, I will focus on providing a summary of findings regarding secondary prevention. However, some evidence for the cost-effectiveness of primary prevention can be found in, for instance, Secker-Walker et al. (1997), Russel (2009) and Owen et al. (2012).

The cost-effectiveness of secondary prevention methods targeting smoking has been extensively studied; see, for instance Woolacott et al. (2002), Keating and Lyseng-Williamson (2010), Hughes (2010) and Bolin (2012). The general finding is that secondary prevention methods are cost-effective. The advances in smoking-cessation treatment have made future smoking-related ill health increasingly avoidable (Reda et al., 2008; Abrams et al., 2010).

Similarly, the evidence for the cost-effectiveness of interventions designed in order to reduce the adverse health effects from alcohol abuse has been reviewed in, for instance, Anderson et al. (2009). They concluded that making alcohol more expensive, restricting alcohol availability

and bans on alcohol advertising, are cost-effective measures for reducing harm. This finding corroborates the findings by, for instance, Lai et al. (2007), who analysed the cost-effectiveness of population-level alcohol control strategies (Estonian setting). They found that taxation and advertising bans were the most cost-effective available policy measures. Cobiac et al. (2009) studied a comprehensive set of both public and individual interventions aiming at preventing harm from alcohol abuse for their cost-effectiveness (Australian setting). All interventions except the residential intervention were highly cost effective. The cost-effectiveness of pharmaceutical treatment for alcohol dependence was studied by Rychlik et al. (2003). They studied the cost-effectiveness of adjuvant treatment with acamprosate (all participants received a psychosocial intervention) and found acamprosate treatment to be cost-effective (about EUR 5 000 lower cost per abstinent person as compared to no acamprosate treatment).

Interventions targeting obesity have been reviewed and evaluated for their cost-effectiveness in several studies, for instance, Ara et al. (2012), Neovius and Narbo (2008), Avenell et al. (2004) and O'Meara et al. (2002). Ara et al. reviewed the evidence for the cost-effectiveness of anti-obesity treatment using orlistat, sibutramine and rimonaban, respectively. They identified and included 94 studies. All three pharmaceutical treatments were found to be cost-effective when using a willingness-to-pay threshold of GBP 20 000 per gained quality-adjusted life year. These results corroborate the earlier findings by Neovius and Narbo (2008) and O'Meara et al. (2002). Avenell et al. (2004) reviewed a broad range of anti-obesity interventions for their long-term cost-effectiveness. They concluded that identifying and targeting high-risk individuals with anti-obesity drug treatment or surgery resulted in a cost per additional quality-adjusted life year of no more than GBP 13 000

The evidence for the cost-effectiveness of internet-based learning in order to improve dietary behaviour was reviewed by Harris et al. (2011). They identified and included 43 studies and concluded that internet-based methods designed for improving dietary behaviour are not cost-effective.

Owen et al. (2012) reviewed the evidence for the cost-effectiveness of public health interventions. They included 21 studies from which UK estimates were inferred. The majority of the interventions targeted smoking, alcohol consumption or physical activity. A majority – 85 per cent – were cost-effective at a GBP 20 000 threshold (15 per cent were cost

saving). Gordon et al. (2007) reviewed the cost-effectiveness of behavioural interventions for smoking, alcohol, diet and physical activity. They concluded that all reported incremental cost-effectiveness ratios were low (for smoking-cessation programmes < EUR 14 000 per quality-adjusted life year gained), in relation to other health interventions. However, as the reported results in the reviewed studies regarding alcohol and dietary vary significantly, no reliable conclusions could be drawn regarding the cost effectiveness of interventions in these areas.

4. Conclusions and discussion

In this paper, I have provided a summary of what is known regarding (1) health risks, and healthcare and productivity costs, and (2) the effectiveness and cost-effectiveness of primary and secondary prevention programmes, associated with smoking, alcohol abuse, nutritional choices, physical activity, and obesity. Health risks associated with smoking are well-established and quantitatively large as compared to other health risks included in this study. Consequently, healthcare costs and costs related to productivity that can be attributed to smoking are fairly well-known, and both primary and secondary effective and cost-effective preventive interventions against smoking are available. The health risks associated with alcohol consumption are also considerable – although not as high as those for daily smokers – for those who consume excessive amounts. The risks decrease with consumption and some studies have even found beneficial health effects associated with moderate consumption. Alcohol-attributable healthcare and productivity costs are also relatively well-known. Some evidence suggests that primary and secondary alcohol prevention may be both effective and cost-effective. However, the evidence is less reliable than the corresponding evidence for smoking.

The picture is less clear when it comes to nutritional choices and physical activity. The main reason for this is that there is conflicting evidence concerning the health risks associated with particular diets. In particular, the predominant view of the health risks associated with dietary saturated fat and high glycemix-index carbohydrates may have to be modified. Moreover, the evidence for the effectiveness and cost-effectiveness of interventions that target behavioural changes regarding diet and physical

activity is generally weak – the exception being interventions designed for preventing the development of obesity in children. The health risk, and healthcare and productivity costs, associated with inadequate physical activity are fairly well-known. However, the healthcare and productivity costs associated with specific dietary patterns, and specific foods, are disputable, due to the unclear relationships between diet and future health outcomes. The cost-effectiveness of interventions designed for changing health behaviours is largely unknown, although some evidence suggests that internet-based interventions in order to change dietary behaviour are not cost-effective.

The health risks, and the healthcare and productivity costs, associated with obesity are well-established. Some evidence suggests that mild obesity may not be detrimental to health, though. Both primary and secondary anti-obesity programmes that target child and adolescent behaviour are potentially effective, although there are unresolved issues as regards the effectiveness of specific interventional components. The cost-effectiveness of pharmaceutical anti-obesity treatments has been systematically reviewed in a number of studies. Drug-based anti-obesity treatments are cost-effective, in particular when high-risk individuals are targeted.

Table 1 summarizes the conclusions reached in this study as regards relative risks, main health effects, the prevalence of risky behaviour, the effects on healthcare and productivity costs and the effectiveness and cost-effectiveness of primary and secondary prevention.

Table 1. Health-related behaviours/conditions

		CONSEQUENCES				
HEALTH BEHAVIOUR		Main health effects	Relative risk (RR)	Prevalence of risky behaviour or condition (high-income EU countries)	Primary prevention (PP)/secondary prevention (SP)	Cost-effectiveness of primary and secondary prevention (PP and SP)
	Smoking	Lung cancer; respiratory diseases; cardiovascular diseases	RR varies between 20 and 2, for lung cancer and ischemic heart disease	Between 15 % and 25 % in the Nordic countries	Some evidence for effectiveness /Established effectiveness	PP: < USD 200 per additional life-year (Secker-Walker et al., 1997); GBP 7 200 per QALY (Owen et al., 2009) SP: estimates in the range: cost saving – ≈ EUR 25 000 per additional QALY (Bolin, 2012)
	Alcohol consumption	Liver disease; cancer	RRs between about 9 for liver cirrhosis and about 4 for cancer of the mouth, oropharynx, and esophagus	About 25 % in high-income countries (EU)	Some evidence for the effectiveness of interventions targeting school children and for public interventions (advertising bans, taxes etc) /established effectiveness of drug-based treatments	PP: estimates in the range: cost saving – ≈ GBP 25 000 (Owen et al., 2009) SP: cost saving - ≈ EUR 24 000 (Owen et al., 2009; Gordon et al., 2007)
	Nutritional choices	Different cancers; cardiovascular diseases; obesity	Dietary fruit and vegetables: RRs between 0.90 and 0.99 for cancer and cardiovascular disease	n.a	Some evidence for the effectiveness of interventions targeting children (nutrition and physical activity) in order to prevent obesity/established effectiveness of both non-pharmaceutical and pharmaceutical based interventions	PP: n.a. SP: n.a.
	Obesity	Different cancers; cardiovascular disease	Obesity: RRs between 1.2 and 1.5 for cancer; RR between 1.4 and 2.6 for cardiovascular disease	23 %		PP: n.a. SP: < GBP 20 000 per additional QALY; < < GBP 13 000 per additional QALY (treating high-risk individuals with anti-obesity drugs)
	Physical activity	Different cancers; cardiovascular diseases; obesity	Physical activity: RRs between 1.5 and 2 for cancer and cardiovascular disease	≈ 70 % inactive or insufficiently active (WHO, 2009; Table A2)		PP: n.a. SP: n.a.

Source: In the table.

Note: Health-related behaviours/conditions: summary of associated health risk, prevalence, the effects on health and healthcare costs, and effectiveness and the cost effectiveness of prevention. Relative risk is defined as the risk associated with the condition faced by an exposed individual in relation to the corresponding risk faced by a non-exposed individual.

4.1 Challenges for future research, institutions and policy-making

This study has identified several instances of inadequacies regarding the information needed for the design of efficient public health policies. In particular, this pertains to the epidemiological evidence regarding the association between health behaviours and health risks, and the knowledge concerning the effectiveness and cost-effectiveness of primary and secondary prevention. The following issues need to be addressed by future research:

- The epidemiological evidence regarding diet and health risks;
- The effectiveness of primary and secondary prevention. As regards smoking, the current knowledge seems sufficient. Interventions aimed at reducing the adverse health effects that result from alcohol abuse, poor diet, and too little physical exercise need to be further explored;
- The cost-effectiveness of interventions in all five areas needs to be further studied. This is the case because published cost-effectiveness studies do not take into account that the implementation of new policy initiatives needs to be financed. Reallocating resources through the tax system will involve efficiency losses that need to be taken into account when calculating cost-effectiveness measures. In principle, bounded rationality should be taken into account when assessing these losses, that is, time and resources may be inefficiently allocated before the tax system is changed or imposed. Moreover, the knowledge of the effectiveness of available policy measures is inadequate in some areas. Thus, the cost-effectiveness of those same measures needs to be re-assessed when new and improved information has been presented.

References

- Abrams, D.B, Graham, A.L, Levy, D.T., Mabry, P.L. and Orleans, C.T. (2010), Boosting population quits through evidence-based cessation treatment and policy, *American Journal of Preventive Medicine* 38, S351-S363.
- Anderson, P., Chisholm, D. and Fuhr, D.C. (2009), Effectiveness and cost-effectiveness of policies and programmes to reduce the harm caused by alcohol, *Lancet* 373, 2234-2246.

- Ara, R., Blake, L., Gray, L., Hernandez, M., Crowther, M. and Dunkley, A. et al. (2012), What is the clinical effectiveness and cost-effectiveness of using drugs in treating obese patients in primary care? A systematic review, *Health Technol Assess* 16.
- Avenell, A., Broom, J., Brown, T.J., Poobalan, A., Aucott, L. and Stearns S.C. et al. (2004), Systematic review of the long-term effects and economic consequences of treatments for obesity and implications for health improvement, *Health Technol Assess* 8.
- Baker, P.R.A., Francis, D.P., Soares, J., Weightman, A.L. and Foster, C. (2011), Community wide interventions for increasing physical activity, *Cochrane Database of Systematic Reviews Issue 4*. Art. No.: CD008366. DOI: 10.1002/14651858.CD008366.pub2.
- Balsa, A.I., Homer, J.F., Fleming, M. and French, M.T. (2008), Alcohol consumption and health among elders, *The Gerontologist* 48, 622-636.
- Becker, G. (1964), *Human Capital*, Columbia University Press (for National Bureau of Economic Research), New York.
- Becker, G. and Murphy, K. (1988), A theory of rational addiction, *Journal of Political Economy* 96, 675-700.
- Bolin, K. (2011), Health Production, in S. Glied and P.C. Smith (eds.), *Oxford Handbook of Health Economics*, Oxford University Press, Oxford.
- Bolin, K. (2012), Economic value of smoking cessation therapies: A critical and systematic review of simulation-generated evidence, *Pharmacoeconomics* 30, 551-564.
- Bolin, K., Borgman, B., Gip, C. and Wilson, K. (2011), Current and future avoidable cost of smoking – Estimates for Sweden, *Health Policy* 103, 83-91.
- Bolin, K., Jacobson, L. and Lindgren, B. (2001), The family as the producer of health – When spouses are Nash bargainers, *Journal of Health Economics* 20, 349-362.
- Bolin, K., Jacobson, L. and Lindgren, B. (2002a), The family as the producer of health – When spouses act strategically, *Journal of Health Economics* 21, 475-495.
- Bolin, K., Jacobson, L. and Lindgren, B. (2002b), Employer investments in employee health – Implications for the family as health producer, *Journal of Health Economics* 21, 563-583.
- Bolin, K. and Lindgren, B. (2006), *Motion – de samhällsekonomiska kostnaderna av inaktivitet och oregelbunden fysisk aktivitet*, Report 2006, Department of Economics, Lund University.
- Bolin, K. and Lindgren, B. (2007), Smoking, healthcare cost, and loss of productivity in Sweden 2001, *Scandinavian Journal of Public Health* 35, 187-196.
- Bolin, K. and Lindgren, B. (2012), The double faceted nature of health investment – Implications for equilibrium and stability in a demand-for-health framework, NBER Working Paper 17789.
- Brinn, M.P., Carson, K.V., Esterman, A.J., Chang, A.B. and Smith, B.J. (2010), Mass media interventions for preventing smoking in young people, *Cochrane Database of Systematic Reviews Issue 11*. Art. No.: CD001006. DOI: 10.1002/14651858.CD001006.pub2.
- Brunner, E., Rees, K., Ward, K., Burke, M. and Thorogood, M. (2007), Dietary advice for reducing cardiovascular risk, *Cochrane Database of Systematic Reviews Issue 4*. Art. No.: CD002128. DOI: 10.1002/14651858.CD002128.pub3.

- Cahill, K., Stead, L-F. and Lancaster, T. (2011), Nicotine receptor partial agonists for smoking cessation, *Cochrane Database of Systematic Reviews Issue 2*. Art. No.: CD006103. DOI: 10.1002/14651858.CD006103.pub5.
- Cawley, J. and Ruhm, C-J. (2011), The economics of risky health behaviors, IZA Discussion Paper 5728, Bonn.
- Chaloupka, F.J., Straif, K. and Leon, M.E. (2011), Effectiveness of tax and price policies in tobacco control, *Tobacco Control* 20, 235-238.
- Civljak, M., Sheikh, A., Stead, L.F. and Car, J. (2010), Internet-based interventions for smoking cessation, *Cochrane Database of Systematic Reviews Issue 9*. Art. No.: CD007078. DOI: 10.1002/14651858.CD007078.pub3.
- Cobiac, L., Vos, T., Doran, C. and Wallace, A. (2009), Cost-effectiveness of interventions to prevent alcohol-related disease and injury in Australia, *Addiction* 104, 1646-1655.
- Curioni, C. and André, C. (2006), Rimonabant for overweight or obesity, *Cochrane Database of Systematic Reviews Issue 4*. Art. No.: CD006162. DOI: 10.1002/14651858.CD006162.pub2.
- Dobbins, M., DeCorby, K., Robeson, P., Husson, H. and Tirilis, D. (2009), School-based physical activity programs for promoting physical activity and fitness in children and adolescents aged 6-18, *Cochrane Database of Systematic Reviews Issue 1*. Art. No.: CD007651. DOI: 10.1002/14651858.CD007651.
- Estaquio, C., Castetbon, K., Emmanuelle, K.G., Bertrais, S., Deschamps, V., Dauchet, L., Péneau, S., Galan, P. and Hercberg, S. (2008), The French national nutrition and health program score is associated with nutritional status and risk of major chronic disease, *Journal of Nutrition* 138, 946-953.
- Fagerström, K. (2002), The epidemiology of smoking. Health consequences and benefits of cessation, *Drugs* 62 (Suppl 2), 1-9.
- Finkelstein, E.A., Brown, D.S., Wraga, L.A., Allaire, B.T. and Hoerger, T.J. (2010), Individual and aggregate years-of-life-lost associated with overweight and obesity, *Obesity* 18, 333-339.
- Forster, M. (2001), The meaning of death: Some simulations of a model of healthy and unhealthy consumption, *Journal of Health Economics* 20, 613-638.
- Foster, C., Hillsdon, M. and Thorogood, M. (2005), Interventions for promoting physical activity, *Cochrane Database of Systematic Reviews Issue 1*. Art. No.: CD003180. DOI: 10.1002/14651858.CD003180.pub2.
- Foxcroft, D.R. and Tsertsvadze, A. (2011), Universal family-based prevention programs for alcohol misuse in young people, *Cochrane Database of Systematic Reviews Issue 9*. Art. No.: CD009308. DOI: 10.1002/14651858.CD009308.
- Fulponi, L. (2009), Policy initiatives concerning diet, health and nutrition, *OECD Food, Agriculture and Fisheries Working Papers* 14, OECD, Paris.
- Garrett, N.A., Brasure, M., Schmitz, K.H., Schultz, M.M. and Huber, M.R. (2004), Physical inactivity: Direct cost to a health plan, *American Journal of Preventive Medicine* 27, 304-309.
- Gaziano, J.M., Buring, J.E., Breslow, J.L., Goldhaber, S.Z., Rosner, B., VanDenburgh, M., Willett, W. and Hennekens, C.H. (1993), Moderate alcohol intake, increased levels of high-density lipoprotein and its subfractions, and decreased risk of myocardial infarction, *New England Journal of Medicine* 329, 1829-1834.

- Gelber, R.P., Gaziano, J.M., Orav, E.J., Manson, J.E., Buring, J.E. and Kurth, T. (2008), Measures of obesity and cardiovascular risk among men and women, *Journal of the American College of Cardiology* 52, 605-615.
- Gordon, L., Graves, N., Hawkes, A. and Eakin, E. (2007), A review of the cost-effectiveness of face-to-face behavioural interventions for smoking, physical activity, diet and alcohol, *Chronic Illness* 3, 101-129.
- Grimshaw, G. and Stanton, A. (2006), Tobacco cessation interventions for young people, *Cochrane Database of Systematic Reviews Issue 4*. Art. No.: CD003289. DOI: 10.1002/14651858.CD003289.pub4.
- Grossman, M. (1972), On the concept of health capital and the demand for health, *Journal of Political Economy* 80, 223-249.
- Grossman, M. (2000), The human capital model, in A.J. Culyer and J.P. Newhouse (eds.), *Handbook of Health Economics* 1A, Elsevier, New York.
- Harris, J., Felix, L., Miners, A., Murray, E., Michie, S. and Ferguson, E. et al. (2011), Adaptive e-learning to improve dietary behaviour: A systematic review and cost-effectiveness analysis, *Health Technology Assessment* 15.
- Hajek, P., Stead, L.F., West, R., Jarvis, M. and Lancaster, T. (2009), Relapse prevention interventions for smoking cessation, *Cochrane Database of Systematic Issue 1*. Art. No.: CD003999. DOI: 10.1002/14651858.CD003999.pub3.
- Hjalte, K., Isacson, S.O., Lindgren, B. and Wilhelmsen, L. (1985), Vad kostar tobaksbrukets medicinska skadeverkningar?, *Läkartidningen* 82, 2978-2981.
- Hooper, L., Summerbell, C.D., Thompson, R., Sills, D., Roberts, F.G., Moore, H. and Smith, D.G. (2011), Reduced or modified dietary fat for preventing cardiovascular disease, *Cochrane Database of Systematic Reviews Issue 7*. Art. No.: CD002137. DOI.
- Hughes, J.R. (2009), How confident should we be that smoking cessation treatments work?, *Addiction* 104, 1637-1640.
- Jarl, J., Johansson, P., Eriksson, A., Eriksson, M., Gerdtham, U.G., Hemström, Ö., Hradilova Selin, K., Lenke, L., Ramstedt, M. and Room, R. (2008), The societal costs of alcohol consumption: An estimation of the economic and human cost in Sweden, 2002, *European Journal of Health Economics* 9, 351-360.
- Keating, G.M. and Lyseng-Williamson, K.A. (2010), Varenicline: A pharmacoeconomic review of its use as an aid to smoking cessation, *Pharmacoeconomics* 28, 231-254.
- Kitahara, C.M. (2010), Low-glycemic load diets: How does the evidence for prevention of disease measure up?, *Journal of the American Dietetic Association* 110, 1818-1819.
- Konnopka, A., Bödemann, M. and König, H.H. (2011), Health burden and costs of obesity and overweight in Germany, *European Journal of Health Economics* 12, 345-352.
- Lai, T., Habicht, J., Reinap, M., Chisholm, D. and Baltussen, R. (2007), Costs, health effects and cost-effectiveness of alcohol and tobacco control strategies in Estonia, *Health Policy* 84, 75-88.
- Laibson, D. (1997), Golden eggs and hyperbolic discounting, *Quarterly Journal of Economics* 112, 443-477.
- Lantz, P.M., Jacobson, P.D., Warner, K.E., Wasserman, J., Pollack, H.A., Berson, J. and Ahlstrom, A. (2000), Investing in youth tobacco control: A review of smoking prevention and control strategies, *Tobacco Control* 9, 47-63.

- Levitan, E.B., Mittleman, M.A., Hakansson, N. and Wolk, A. (2007), Dietary glyce-
mic index, dietary glyce-
mic load, and cardiovascular disease in middle aged and
older Swedish men, *American Journal of Clinical Nutrition* 85, 1521-1526.
- Levitan, E.B., Mittleman, M. and Wolk, A. (2009), Dietary glyce-
mic index, dietary
glyce-
mic load and mortality among men with established cardiovascular disease,
European Journal of Clinical Nutrition 63, 552-557.
- Levy, D.T., Graham, A.L., Mabry, P.L., Abrams, D.B. and Orleans, C.T. (2010),
Modeling the impact of smoking-cessation treatment policies on quit rates, *Ameri-
can Journal of Preventive Medicine* 38, S364-S372
- Liang, W. and Chikritzhs, T. (2010), Reduction in alcohol consumption and health
status, *Addiction* 106, 75-81.
- Liljas, B. (1998), The demand for health with uncertainty and insurance, *Journal of
Health Economics* 17, 153-170.
- Liljas, B. (2000), Insurance and imperfect financial markets in Grossman's demand
for health model – A reply to Tabata and Ohkusa, *Journal of Health Economics* 19,
811-820.
- Lock, K., Pomerleau, J., Causer, L. and McKee, M. (2004), Low fruit and vegetable
consumption, in M. Ezzati et al. (eds.), *Comparative Quantification of Health Risks:
Global and Regional Burden of Disease Attributable to Selected Major Risk Fac-
tors*, World Health Organization, Geneva.
- Lundborg, P. (2007), Does smoking increase sick-leaves? Evidence using register data
on Swedish workers, *Tobacco Control* 16, 114-118.
- Mente, A., de Koning, L., Shannon, H.S. and Anand, S.S. (2009), A systematic re-
view of the evidence supporting a causal link between dietary factors and coronary
heart disease, *Archives of Internal Medicine* 169, 659-669.
- Moreira, M.T., Smith, L.A. and Foxcroft, D. (2009), Social norms interventions to
reduce alcohol misuse in university or college students, *Cochrane Database of Sy-
stematic Reviews Issue 3*. Art. No.: CD006748. DOI:
10.1002/14651858.CD006748.pub2.
- Mozaffarian, D. (2011), The great fat debate: Taking the focus off of saturated fat,
Journal of the American Dietetic Association 111, 665-666.
- Neovius, M. and Narbro, K. (2008), Cost-effectiveness of pharmacological anti-
obesity treatments: A systematic review, *International Journal of Obesity* 32, 1752-
1763.
- Ni, M.C., Rodgers, A., Pan, W.H., Gu, D.F. and Woodward, M. (2004), Body mass
index and cardiovascular disease in the Asia-Pacific Region: An overview of 33
cohorts involving 310 000 participants, *International Journal of Epidemiology* 33,
751-758.
- O'Meara, S., Riemsma, R., Shirran, L., Mather, L. and ter Riet, G. (2002), The clini-
cal effectiveness and cost effectiveness of sibutramine in the management of obesi-
ty: A technology assessment, *Health Technology Assessment* 6.
- Oude Luttikhuis, H., Baur, L., Jansen, H., Shrewsbury, V.A., O'Malley, C., Stolk,
R.P. and Summerbell, C.D. (2009), Interventions for treating obesity in children,
Cochrane Database of Systematic Issue 1. Art. No.: CD001872. DOI:
10.1002/14651858.CD001872.pub2.
- Owen, L., Morgan, A., Fischer, A., Ellis, S., Hoy, A. and Kelly, M.P. (2012), The
cost-effectiveness of public health interventions, *Journal of Public Health* 34, 37-45.

- Owen, L., Morgan, A., Fischer, A., Ellis, S., Hoy, A. and Kelly, M.P. (2012), The cost-effectiveness of public health interventions, forthcoming in *Journal of Public Health*.
- Padwal, R.S., Rucker, D., Li, S.K., Curioni, C. and Lau, D.C.W. (2003), Long-term pharmacotherapy for obesity and overweight, *Cochrane Database of Systematic Reviews* Issue 4. Art. No.: CD004094. DOI: 10.1002/14651858.CD004094.pub2.
- Rasmussen, S.R., Prescott, E., Sørensen, T.I.A. and Sjøgaard, J. (2004), The total life time costs of smoking, *European Journal of Public Health* 14, 95-100.
- Reda, A.A., Kaper, J. and Fikretler, H. et al. (2008), Healthcare financing systems for increasing the use of tobacco dependence treatment, *Cochrane Database of Systematic Reviews* Issue 4: CD004305.
- Rehm, J., Gmel, G., Sempos, C.T. and Trevisan, M. (2003), Alcohol-related morbidity and mortality, *Alcohol Research and Health* 27, 39-51.
- Rehm, J., Mathers, C., Popova, S., Thavorncharoensap, M., Teerawattananon, Y. and Patra, J. (2009), Global burden of disease and injury and economic cost attributable to alcohol use and alcohol-use disorders, *Lancet* 373, 2223-2233.
- Renehan, A.G., Tyson, M., Egger, M., Heller, R.F. and Zwahlen, M. (2008), Body-mass index and incidence of cancer: A systematic review and meta-analysis of prospective observational studies, *Lancet* 371, 569-578.
- Roberts, S.B. (2000), High-glycemic index foods, hunger, and obesity: Is there a connection?, *Nutrition Reviews* 58, 163-169.
- Rychlik, R., Siedentop, H., Pfeil, T. and Daniel, D. (2003), Cost-effectiveness of adjuvant treatment with acamprostate in maintaining abstinence in alcohol dependent patients, *European Addiction Research* 9, 59-64.
- Roberts, L. and Lindgren, B. (2001), Do smokers pay for their sin? Utilisation of the social welfare system and distribution of costs and benefits among smokers and non-smokers in Sweden, Lund University Centre for Health Economics (LUCHE), Lund.
- Royal College of Physicians (1962), *Smoking and Health, Summary and Report of the Royal College of Physicians of London on Smoking in Relation to Cancer of the Lung and Other Diseases*, Pitman Publishing, New York.
- Russel, L.B. (2009), Preventing chronic disease: An important investment, but don't count on cost savings, *Health Affairs* 28, 42-45.
- Rösner, S., Hackl-Herrwerth, A., Leucht, S., Lehert, P., Vecchi, S. and Soyka, M. (2010a), Acamprostate for alcohol dependence, *Cochrane Database of Systematic Reviews* Issue 9. Art. No.: CD004332. DOI: 10.1002/14651858.CD004332.pub2.
- Rösner, S., Hackl-Herrwerth, A., Leucht, S., Vecchi, S., Srisurapanont, M. and Soyka, M. (2010b), Opioid antagonists for alcohol dependence, *Cochrane Database of Systematic Reviews* Issue 12. Art. No.: CD001867. DOI: 10.1002/14651858.CD001867.pub3.
- Secker-Walker, R.H., Worden, J.K., Holland, R.R., Flynn, B.S. and Detsky, A.S. (1997), A mass media programme to prevent smoking among adolescents: Costs and cost-effectiveness, *Tobacco Control* 6, 207-212.
- Shafey O., Dolwick, S. and Guindon, G.E. (eds.) (2003), *Tobacco Control Country Profiles, second edition*, The 12th World Conference on Tobacco or Health, American Cancer Society, Atlanta, GA.

- Siri-Tarino, P.W., Sun, Q., Hu, F.B. and Krauss, R.M. (2010a), Meta-analysis of prospective cohort studies evaluating the association of saturated fat with cardiovascular disease, *American Journal of Clinical Nutrition* 91, 535-545.
- Siri-Tarino, P.W., Sun, Q., Hu, F.B. and Krauss, R.M. (2010b), Saturated fat, carbohydrate, and cardiovascular disease, *American Journal of Clinical Nutrition* 91, 502-509.
- Stead, L.F., Perera, R., Bullen, C., Mant, D. and Lancaster, T. (2008), Nicotine replacement therapy for smoking cessation, *Cochrane Database of Systematic Reviews* Issue 1, Art. No.: CD000146. DOI: 10.1002/14651858.CD000146.pub3.
- Sturm, R. (2002), The effects of obesity, smoking, and drinking on medical problems and costs, *Health Affairs* 21, 245-253.
- Takachi, R.I., Inoue, M., Ishihara, J., Kurahashi, N., Iwasaki, M., Sasazuki, S., Iso, H., Tsubono, Y. and Tsugan, S. (2008), Fruit and vegetable intake and risk of total cancer and cardiovascular disease: Japan Public Health Center-Based Prospective Study, *American Journal of Epidemiology* 167, 59-70.
- Thomas, D., Elliott, E.J. and Baur, L. (2007), Low glycaemic index or low glycaemic load diets for overweight and obesity, *Cochrane Database of Systematic Reviews* Issue 3. Art. No.: CD005105. DOI: 10.1002/14651858.CD005105.pub2
- Thomas, R.E. and Perera, R. (2006), School-based programmes for preventing smoking, *Cochrane Database of Systematic Reviews* Issue 3. Art. No.: CD001293. DOI: 10.1002/14651858.CD001293.pub2.
- Trogdon, J.G., Finkelstein, E.A., Hylands, T., Dellea, P.S. and Kamal-Bahl, S.J. (2008), Indirect costs of obesity: A review of the current literature, *Obesity Reviews* 9, 489-500.
- Thun, M.J., Apicella, L.F. and Henley, S.J. (2000), Smoking vs other risk factors as the cause of smoking-attributable deaths. Confounding in the courtroom, *JAMA* 284, 706-712.
- Thun, M.J., Peto, R., Lopez, A.D., Monaco, J.H., Henley, S.J., Heath, C.W. and Doll, R. (1997), Alcohol consumption and mortality among middle-aged and elderly U.S. adults, *New England Journal of Medicine* 337, 1705-1714.
- Wakefield, M. and Chaloupka, F. (2000), Effectiveness of comprehensive tobacco control programmes in reducing teenage smoking in the USA, *Tobacco Control* 9, 177-186.
- Waters, E., de Silva-Sanigorski, A., Hall, B.J., Brown, T., Campbell, K.J., Gao, Y., Armstrong, R., Prosser, L. and Summerbell, C.D. (2011), Interventions for preventing obesity in children, *Cochrane Database of Systematic Reviews* Issue 12. Art. No.: CD001871. DOI:10.1002/14651858.CD001871.pub3.
- Withrow, D. and Alter, D.A. (2011), The economic burden of obesity worldwide: A systematic review of the direct costs of obesity, *Obesity Reviews* 12, 131-141.
- Woolcott, N.F., Jones, L., Forbes, C.A. et al. (2002), The clinical effectiveness and cost-effectiveness of bupropion and nicotine replacement therapy for smoking cessation: A systematic review and economic evaluation, *Health Technology Assessment* 6.
- World Cancer Research Fund (2009), Policy and Action for Cancer Prevention. Food, Nutrition, and Physical Activity: A Global Perspective, American Institute for Cancer Research, Washington DC.
- World Health Organization (2009), Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks, WHO Press, Geneva.

Zohrabian, A. and Philipson, T.J. (2010), External costs of risky health behaviors associated with leading actual causes of death in the U.S.: A review of the evidence and implications for future research, *International Journal of Environmental Research and Public Health* 7, 2460-2472.

Comment on Bolin: Lifestyle, health and costs – What does available evidence suggest?

Tinna Laufey Ásgeirsdóttir*

Kristian Bolin's paper reviews the available evidence on major health related behaviors. He explains the development of those behaviors over time, the economic theory underlying their examination, their effect on health and costs, the effectiveness of primary and secondary prevention, as well as the cost effectiveness of selected interventions. With such a broad-ranging task, the author needs to pick and choose the included literature carefully. In a comment such as this, the easiest route to take would probably be to pick out papers which might have been justly included. However, in this case, I think that would not be very helpful. No two people would include exactly the same papers in such an overview and the available evidence included is well selected and provides a balanced picture of the literature.

Economics is about how individuals and societies maximize their well being (utility) given the scarce resources available, shedding some light on welfare that is gained or lost by government actions. In the discussion of Kristian Bolin's paper, I will emphasize this point as I think it may have suffered a bit under the broad scope of an ambitious paper. Especially due to the broad scope of the paper, I think it might have given the paper a greater focus to limit the discussion on topics that are not of core importance in this context. This includes human-capital theory and the demand for health, which provide limited scientific or policy guidance on

* Department of Economics, University of Iceland, ta@hi.is.

behavioral interventions. The same goes for other explanations, for example on randomization, natural experiments and attributable fractions, although not to the same extent.

According to basic economic theory, social well-being is maximized when individuals have freedom to actions that concern them and their associates, as long as no market failures are present. In the presence of market failures, governments can increase societal welfare. Economically, the aim is thus not for the public to be of optimal weight (e.g. $18.5 < \text{BMI} < 25$), smoke the least or consume alcohol consistent with public-health recommendations. The final aim is not for people to be the healthiest, but to maximize their overall well-being. Therefore, the main market failures that should be taken into consideration regarding body weight, smoking, and alcohol consumption are vastly important when considering the economic efficiency of government action. The action taken needs to be in accordance with the market failure at hand, for example, so that it affects the individuals whose behavior deviates from the efficient amount, and not that of others. In my mind, when aiming at alleviating behavioral inefficiencies, it is more helpful to use classifications of market failures and related interventions, rather than using classifications from health sciences, such as primary and secondary prevention.

The core of the matter is this: From an economic perspective, there are two factors that need to be considered in government policy making. First, it is necessary to assess whether there are sound economic reasons for government intervention – the market failures. If that is the case, the cost effectiveness of market-failure-appropriate interventions can be evaluated.

The government possesses various controls, such as regulations, directives, provision of information and subsidies. If the government chooses to correct a market failure by intervening, it also needs to be able to give an account of what method should be used so that the benefits will outweigh the costs. Kristian Bolin gives an overview of the literature that compares the benefits and costs of many such interventions. However, the first step of theoretical justifications for such intervention is important. Such justifications are certainly mentioned in Kristian Bolin's paper, but not in a very systematic way, and the interventions are not discussed in their light. It is not enough for an action to be cost effective for the government to intervene. Fortunately, most decisions made by individuals

and businesses alike are cost effective. That is why those actions are taken. That does not in and of itself justify an intervention. Different market failures call for different actions. It is, for example, unlikely that information can efficiently change the consumption inefficiencies due to externalities. The provision of information would be more suited to mend a market failure related to information problems. For this reason, I would like to focus a bit on those theoretical issues related to the lifestyle behaviors of interest.

1. Externalities of obesity, smoking and alcohol abuse

Costs that affect a third party and move behavior away from optimal social well-being need to be taken into account. Health behavior often leads to such external costs, as is mentioned by Kristian Bolin. The two main factors of “externalities” due to, for example, obesity are increased health expenditures (Borg et al., 2005; Sampalis et al., 2004; Finkelstein et al., 2003) and benefits for those who do not work because of their body composition (Ásgeirsdóttir, 2011). However, this inefficiency exists because of government intervention, namely public financing of the health-care system and income replacements. It could thus be argued that a way of getting rid of these “externalities” would be to abolish those insurance systems. Thus, strictly speaking, this case is not an example of externalities, but of moral hazard, brought about by insurance. Since those arrangements are so deeply rooted and widely accepted in the Nordic welfare mentality, it may be justifiable to refer to them as externalities.

As a counterbalance to the negative externalities caused by an increased use of health-care services, the positive externalities of pension fund savings on other tax payers need to be considered. This would result in lower pensions for those who are overweight, smoke, or consume excessive amounts of alcohol, despite them contributing on the same terms as others. Working age individuals account for a large part of premature deaths due to obesity and therefore cause output losses. However, this does not affect other citizens since wage compensations are obviously not paid in these instances. On the contrary, the pension can be saved because of the shorter chronological age of these individuals. “Savings” due to premature deaths do possibly somewhat counterbalance an increased cost

in the health-care system due to health behaviors. Furthermore, increased medical service costs in later years can be pointed out, although this is a more complicated issue. This increase can be difficult to estimate. The reason is the difficulty in separating medical costs due to biological age vs. proximity to death.

When externalities are present, price changes through taxation and subsidies are a traditional remedy. Negative externalities can be remedied with taxes and if the market failure involves positive externalities, as is possible with various physical activities, governments can encourage consumption with subsidies or direct provisions. Moreover, regulations or directives can be used to encourage consumption that involves positive externalities, but both the costs and benefits of such actions need to be weighed.

2. Bounds of rationality and internalities

The vast number of diet products that people purchase, along with the ways individuals addicted to tobacco and alcohol try to control their consumption, indicate that self-control problems are present. Self-control has been explained as time inconsistency in discounting. Such an inconsistency can cause an individual not to behave in a manner that maximizes long-term welfare. Suppose that an individual is asked if he wants a beer today or five Euros tomorrow and answers that he prefers the beer. If this individual's options were to get the beer in 50 days or five Euros in 51 days, he might answer that he rather wanted the five Euros. This indicates a higher discount rate if the desiderata are to be enjoyed soon. This has sometimes been termed internalities. Internalities can be thought of as externalities that the individual experiences in the future due to his own decisions. However, the individual did not fully take these future effects into account when originally making the decision. O'Donohue and Rabin (1999) discuss the effect of immediate and delayed rewards when time inconsistency in an individual's discounting is present. And a formal presentation of such discounting can be found in Harris and Laibson (2001). It is difficult to predict what fraction of people discounts in this time-inconsistent way. However, it is important to reflect on how large this group is because actions increasing the well-being of this group

might reduce the welfare of others who do not have this problem – such as if their accessibility to prepared foods or alcohol would be reduced in order to separate the time of consumption from the time of decision making about that consumption. That is, welfare losses can occur if actions are not such that they specifically target this group, and not others.

The concept of anti-market has been used for a market that offers a good in order to reduce the consumption of another good. One traditional, although not widely used example would be the drug Antabus, which makes people very sick if they drink alcohol. In that way, a rational individual can put restrictions on himself by taking measures that make him avoid temptations in the future. Today, obesity surgery seems to be among the treatments that help individuals the most in losing weight in the long run (Sjöström et al., 2003). With better surgery technology, individuals have lost up to 34-40 percent of their overall body weight or 70-80 percent of their excess weight, where up to 86 percent of the individuals suffering from diabetes fully or partly get rid of the disease, 70 percent of too high cholesterol, 79 percent of hypertension, and 84 percent of sleep apnea (Brolin, 2002; Buchwald et al., 2004; Jones, 2000; Leifsson and Gíslason, 2005; Munn et al., 2001; Pories et al., 1995; Zingmond et al., 2005). Several other examples of anti markets could be mentioned, but individuals are known to control their own behavior in various ways. People can, for example, try to escape temptations by going to a sanitarium. Others have gone as far as wiring their teeth together to limit their consumption. In addition, there are 12-step programs, health programs with friends or other group therapies. The existence of those commitment devices, as such measures have sometimes been termed, suggests the presence of self-control issues. However, even in the existence of self-control problems, one needs to ask if anti markets do not work efficiently to solve this. If there is reason to believe that they do not, then inefficiencies in such markets might lead to sensible intervention criteria.

3. Children and insurance-market failure

While the health symptoms of the behaviors in question have been found in children and adolescents, immediate externalities in terms of service burdens at health institutions are very small. However, it cannot be over-

looked that there is a strong association between childhood circumstances and later-in-life outcomes in this regard. Viner and Cole (2005) show, for example, that 52 percent of the British children who were obese at the age of 10 were also obese at the age of 30. Research has similarly shown that children who exercise regularly are more likely to continue that lifestyle in adulthood. The same can be said about complications from smoking and excessive alcohol consumption. The diseases that develop due to this consumption emerge later in life, apart from the addiction that can develop so quickly that individuals may have difficulty quitting when they finally reach the maturity to rightly evaluate the costs and benefits of their consumption. Smoking and alcohol consumption generally start when an individual's ability to make informed decisions regarding own welfare is limited. The same goes for the lifestyle that leads to overweight and obesity. These circumstances are formed by the children's environment and circumstances.

It is clear that it takes time for children to gain the ability needed for such decision making and the factors being discussed here are largely formed in childhood. The perfect market assumption of full rationality does thus not apply to childhood. However, since every individual has an agent, the limited ability to make informed decisions may not be a problem. Parents are meant to make decisions for their children. This is not unlike various circumstances where better informed agents are called upon to make decisions for their principals. In many other instances, agents act on behalf of adult principals regarding their health. Other factors are also involved, however.

Every individual's life is a string of coincidences. The uncertainty that comes with these coincidences is generally not a problem, since it is usually possible to buy certainty through insurance markets. However, that does not go for the most fateful coincidence in every individual's life – the circumstances into which one is born. This involves a considerable risk, such as the possibility that an individual will suffer from obesity in adulthood, or be addicted to tobacco and alcohol. Every individual's life has a determined beginning and therefore he/she cannot buy insurance against this risk due to the simple fact that he is not alive. However, a risk-averse individual facing such a risk would most likely buy insurance. This may be looked at as a market failure in the insurance market. The market failure is due to the timeframe of every individual's life. Life

begins at a certain point in time and before that time, we cannot buy insurance against being born into difficult circumstances.¹

One solution to insurance problems at birth is government insurance, where the government acts so that these individuals – children – are insured against this risk and while it is not possible to provide everyone with exactly the same health, it may also be kept in mind that risk-averse people do not necessarily fully insure. Furthermore, if the aim of the government is to redistribute goods at the beginning of life, the use of monetary or in-kind transfers could be considered. An income transfer gives a family an opportunity to spend the money in a way that they prefer and in most cases therefore leads to a more favorable result from a utility maximization viewpoint. According to economic theory, it is traditionally possible to increase welfare more using money transfers. This suboptimality of in-kind transfers is due to the fact that a transfer that is not in the form of money (in-kind transfer) does not give the recipient an opportunity to choose the consumption combination that provides the most utility. However, in the light of the market failure involved here – a broken insurance market at the beginning of life – it is possible to argue for in-kind subsidies. The reason is that the child is partly being insured against his/her own contextual environment, including parents, who would otherwise have to act as the child's agent in spending the monetary transfers. Thus, many possible interventions, such as school based programs in Scandinavian school systems, could be seen as tackling this market failure, while other interventions do not.

Thus, in short: I think it would be very interesting in future work to separate the analysis of possible behavioral interventions according to the associated market failures and evaluate them according to the severity of the associated inefficiencies brought about by the market failures.

¹ Although the economic reasoning of this article involves a discussion based on market failure assumptions, this is not inconsistent with the theories of fairness in economics pioneered by Kolm (1972). His results show that circumstances are fair when two parties trade in a free market, as long as their opportunities were equal to begin with. In other words, the result would be fair if this uncertainty were eliminated. This goes e.g. for health and health behavior that are parts of what individuals take along with them into life. The above mentioned problem could be looked at from the perspective of a risk-averse individual that makes decisions under John Rawls' veil of ignorance.

References

- Ásgeirsdóttir, T. (2011), Do body weight and gender shape the work force? The case of Iceland, *Economics and Human Biology* 9, 148-156.
- Borg, S., Persson U., Odegaard K., Berglund G., Nilsson, J.A. and Nilsson, P.M. (2005), Obesity, survival, and hospital costs-findings from a screening project in Sweden, *Value Health* 8, 562-571.
- Brolin, R.E. (2002), Bariatric surgery and long-term control of morbid obesity, *Journal of the American Medical Association* 288, 2793-2796.
- Buchwald, H., Avidor, Y., Braunwald, E., Jensen, M.D., Pories, W., Fahrbach, K. and Schoelles, K. (2004), Bariatric surgery: A systematic review and meta-analysis, *Journal of the American Medical Association* 292, 1724-1737.
- Finkelstein, E.A., Fiebelkorn, I.C. and Wang, G.J. (2003), National medical spending attributable to overweight and obesity: How much, and who's paying?, *Health Affairs Suppl W3*, 219-226.
- Harris, C. and Laibson, D. (2001), Dynamic choices of hyperbolic consumers, *Econometrica* 69, 935-957.
- Jones, K.B. (2000), Experience with the Roux-en-Y gastric bypass, and commentary on current trends, *Obesity Surgery* 10, 183-185.
- Leifsson, B.G. and Gíslason, H. (2005), Laparoscopic Roux-en-Y gastric bypass with 2-metre long biliopancreatic limb for morbid obesity: Technique and experience with the 150 first patients, *Obesity Surgery* 15, 25-42.
- Munn, E.C., Blackburn, G.L. and Matthews, J.B. (2001), Current status of medical and surgical therapy for obesity, *Gastroenterology* 120, 669-681.
- O'Donohue, T. and Rabin, M. (1999), Doing it now or later, *American Economic Review* 89, 103-124.
- Pories, W.J., Swanson, M.S., Macdonald, K.G., Long, S.B., Morris, P.G., Brown, B.M., Barakat, H.A., Deramon, R.A., Israel, G., Dolezal, J.M. and Dohm, L. (1995), Who would have thought it? An operation proves to be the most effective therapy for adult-onset diabetes mellitus, *Annals of Surgery* 222, 339-352.
- Sampalis, J.S., Liberman, M., Auger, S. and Christou, N.V. (2004), The impact of weight reduction surgery on health-care costs in morbidly obese patients, *Obesity Surgery* 14, 939-947.
- Sjöström, L., Lindroos, A.K., Peltonen, M., Torgerson, J., Bouchard, C., Carlsson, B., Dahlgren, S., Larsson, B., Narbro, K., Sjöström, C.D., Sullivan, M. and Wedel, H. (2003), Lifestyle, diabetes, and cardiovascular risk factors 10 years after bariatric surgery, *The New England Journal of Medicine* 351, 2683-2693.
- Viner, R.M. and Cole, C. (2005), Adult socioeconomic, educational, social and psychological outcomes of childhood obesity: A national birth cohort study, *British Medical Journal* 330, 1354-1357.
- Zingmond, D.S., McGory, M.L. and Ko, C.Y. (2005), Hospitalization before and after gastric bypass surgery, *Journal of the American Medical Association* 294, 1918-1924.

The economics of long-term care: A survey^{*}

Helmuth Cremer^{**}, Pierre Pestieau^{***} and Gregory Ponthiere^{****}

Summary

This paper surveys recent theoretical economic research on long-term care (LTC). LTC differs from health care: it is about nursing; it is mostly provided by unpaid caregivers (mainly spouses and children), whereas both the market and the state play a modest role. The future of LTC appears to be gloomy: sustained population ageing and recent societal trends (e.g., children's mobility, changes in family values) generate a mounting demand on the state and on the market to provide alternatives to the family. In this paper, we review these causes, and the extent to which we can expect them to fade away in the future. Then, we turn to the design of a sustainable public LTC scheme integrating both the market and the family.

Keywords: long-term care, social insurance, dependence, family solidarity.

JEL classification numbers: I11, I12, I18, J14.

^{*} Financial support from the Chaire "Marché des risques et création de valeur" of the FdR/SCOR is gratefully acknowledged. We would like to thank Erik Schokkaert, Chiara Canta, Thorolfur Matthiasson and the editors for insightful remarks and helpful suggestions.

^{**} Toulouse School of Economics (IDEI, GREMAQ and Iuf), helmuth.cremer@tse-fr.eu.

^{***} CORE, University of Louvain; CREPP, University of Liège; IDEI, Toulouse School of Economics, p.pestieau@ulg.ac.be.

^{****} Ecole Normale Supérieure, Paris – Paris School of Economics, gregory.ponthiere@ens.fr.

Long-term care (LTC) concerns people who depend on help to carry out daily activities such as eating, bathing, dressing, going to bed, getting up or using the toilet. It is delivered informally by families – mainly spouses, daughters and step-daughters – and, to a lesser extent, formally by care assistants, who are paid under some form of employment contract. Formal care is given at home or in an institution (such as care centers and nursing homes). The governments of most industrialized countries are involved in either the provision or financing of LTC services, or often both, although the extent and nature of their involvement differs widely across countries.

In the future, the demand for formal LTC services by the population is likely to grow substantially. LTC needs start to rise exponentially from around the age of 80. The number of individuals aged 80 years and above is growing faster than any other segment of the population. As a consequence, the number of dependent elderly at the European level (EU-27) is expected to grow from about 21 million people in 2007 to about 44 million in 2060 (EC, 2009). Thus, we anticipate an increasing pressure on the resources demanded to provide LTC services for the frail elderly, and this pressure will be on the three institutions currently financing and providing LTC services: the state, the market and the family.¹

These three institutions have their pluses and minuses. The family provides services that are warm, cheap and distortionless. However, these services are restricted to each individual's family circle. Furthermore, some families are very poor and some dependent persons cannot count on family solidarity at all.² The state is the only institution that is universal and redistributive, but its information is quite often limited and its means of financing are distortionary. Finally, the market can be expensive, particularly where it is thin and without public intervention, it only provides services to those who can afford it.³

¹ The existence of LTC provision outside markets leads to ambiguous predictions about the future growth of the formal LTC market. Indeed, as argued by Lakdawalla and Philipson (2002), aging may actually *reduce* the per capita demand for market LTC, provided that it raises the supply of non-market care produced by other elderly people.

² See Duée et al. (2005) on the predicted rise in the number of dependent elderly without family help in France. Note that large divorce rates may substantially increase the role of children in comparison to spouses for LTC provision.

³ Regarding LTC provision in institutions, the quality of LTC services is strongly variable with the LTC techniques used (more or less labor intensive), which significantly affect the health of the elderly (see Cawley et al., 2004).

In assessing the adequacy of LTC financing and provision and in making projections, it is important to bear in mind the extent to which countries will be able to rely on the informal provision of care in the future. The bulk of LTC is indeed provided informally.⁴ Informal provision has no direct bearing on public finances,⁵ but it is not clear that such a situation is desirable and, in any case, will last. Family solidarity is very uneven, and its propensity to provide care could diminish, due to changes in family structure and the growing participation of women in the labor market, which may constrain the future supply of informal care provision within households.

The market for LTC insurance is still negligible, with the exceptions of France and the US. As to the public sector, few countries have a formal social LTC insurance. Even though they do not have a formal social insurance on LTC, most countries devote resources to the financing of LTC services, most often at the local level, but the share of GDP devoted to these is small. One may hope that both private and social LTC insurance will grow substantially in coming decades. But there is a number of problems that both the state and the market have to solve before they can replace family solidarity.

In the next section, we first study the nature of dependency in old age. Then, we present some recent forecasts regarding future needs of LTC. In Section 2, we study what explains the underdevelopment of private insurance for LTC; this is labeled as the LTC insurance market puzzle. Section 3 is devoted to the role of family, more specifically, to different ways of modeling the interactions among parents, spouses and children. These interactions can be triggered by some sort of altruism or just by a mechanism of intergenerational exchanges. Section 4 deals with the design of a sustainable public LTC scheme that integrates the role of the market and the family. A final section concludes the paper.⁶

⁴ According to Norton (2000), about two thirds of LTC is provided informally. Naturally, that figure is a simplification, since there exist strong international differences in LTC provision (see below).

⁵ It clearly has indirect incidence by reducing female labor participation.

⁶ For earlier surveys, see Norton (2000), Brown and Finkelstein (2009), Cremer et al. (2009) and Cremer and Pestieau (2010).

1. Concepts and facts

Loss of autonomy or old-age dependency can be defined as the inability, due to old age, to carry out basic daily activities, such as, for instance, eating, dressing, washing, walking, etc. As a consequence of that inability, dependent elderly people require LTC assistance. The LTC phenomenon is thus permanent, non-accidental and due to old age and, as such, it should be distinguished from other phenomena, such as illness, disability and handicap.

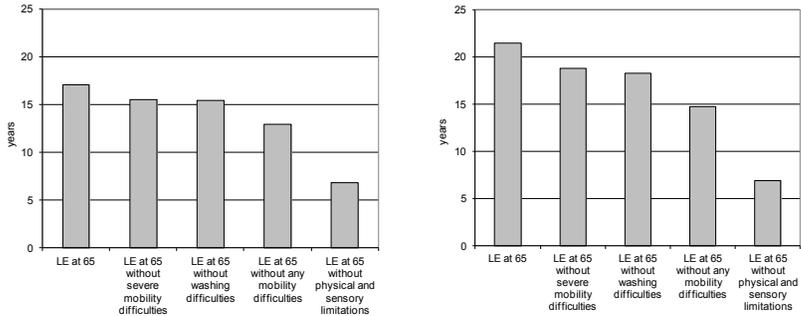
Measuring the loss of autonomy among the elderly is not straightforward, since it requires classifying the elderly as either autonomous or non-autonomous, whereas autonomy is a matter of degree. There exists a *continuum* of health states between the perfectly autonomous young adult and the fully dependent very old person. Several measures have been developed around the world to measure the prevalence of old-age dependency. The well-known Katz scale, which is used by US insurers, counts as dependent the elderly who are unable to carry out at least two out of six activities of daily life (bathing, dressing, transferring, toilet use, eating and continence). Another scale is the AGGIR scale, which is used by the French national system for personal LTC allowances.⁷

The prevalence of various autonomy restrictions at old age is well illustrated by the concept of disability-free life expectancy, i.e. the average life duration without particular autonomy limitations.⁸ Figure 1 shows, in the case of France, the life expectancy at age 65 for men and women, as well as some disability-free life expectancy statistics, for 2003. Obviously, women live, on average, longer than men, and they also enjoy longer periods of life without disability. However, the periods with disabilities are also longer for women than for men. For instance, men spend, on average, 1.5 years with serious mobility restrictions, against 2.7 years for women.

⁷ On the variety of measures of autonomy loss, see Kessler (2007).

⁸ That measure is directly computed on basis of probabilities of different limitations at each age of life. See Pérès et al. (2005) and Cambois et al. (2008).

Figure 1. Life expectancy without disability for men (left) and women (right) France, 2003



Source: Cambois et al. (2008, p. 293).

While those figures give us an idea of the prevalence of the LTC phenomenon, they only concern an economy at a given point in time. One may also be interested in future LTC prevalence. Forecasting future needs is a daring but necessary undertaking. It requires two steps. First one wants to know the relative number of dependent elderly in the future. Second, one has to allocate those individuals among the various types of LTC: formal *versus* informal, private *versus* public. For the first step, we have good forecasts of the future population structure. According to the population projection by main age groups for EU27, the old-age dependency ratio, calculated as the ratio of people aged 65+ relative to the working-age population, will go from 25.4 percent to 53.3 percent over the period 2008-2060.⁹ The dependency ratio of the oldest-old (people aged 80+ over the working-age population) will increase from 6.5 percent to 22 percent over the same period. Dependency does indeed increase with age, particularly after 75; it is more prevalent among women than among men, as shown in Figure 1.

Regarding the proportion of dependent among the elderly, several forecasts can be made, depending on the predicted future evolution of dependency. As discussed by the EC (2009), two broad scenarios can be

⁹ See EC (2009).

used.¹⁰ On the one hand, the “pure demographic” scenario, according to which currently observed age-specific dependency rates will prevail in the future. In other words, there would be no improvement in the dependency status of the elderly population even though average longevity increases.¹¹ On the other hand, the “constant disability” scenario, according to which the duration of the dependency period will remain unchanged in the future, despite the lengthening of life. These two scenarios are illustrated below.

Forecasts under the “pure demographic” scenario are frightening: in EU27, the total number of dependent elderly will, under that scenario, grow from about 20 million in 2007 to 44 million in 2060, which corresponds to a total growth of 115 percent.¹² The number of dependent old individuals will more than double.

Arguably, this “pure demographic” scenario is quite pessimistic, since it assumes that the average lifetime consumption of long-term care services will increase over time. This “pure demographic” scenario also contradicts empirical studies predicting that the duration of the dependency period remains roughly constant in spite of an increase in the average duration of life; see, for instance, Cambois et al. (2008) who study disability in France above the age of 65. Therefore, one may prefer the “constant disability” scenario, in which there is no extension of the morbidity period (in absolute terms) as the total length of life increases. But even under that more optimistic scenario, the number of dependents will still grow substantially in the next decades.¹³

LTC is provided in different settings: formally and informally (some persons receive no care at all). In the case of formal care, it can be at home or in various types of institutions, including nursing homes and long-stay hospitals. Assuming the “pure demographic” scenario, that is, assuming that the probability (at any given age) of receiving formal care at home and formal care in an institution remains constant at the 2007 level, the percentage change in the number of dependent receiving care in an institution would be 185 in EU27 (155 for EU10); for those receiving

¹⁰ In the projection of EC (2009), the dependency rates are drawn from SHARE – Survey on Health, Aging and Retirement in Europe.

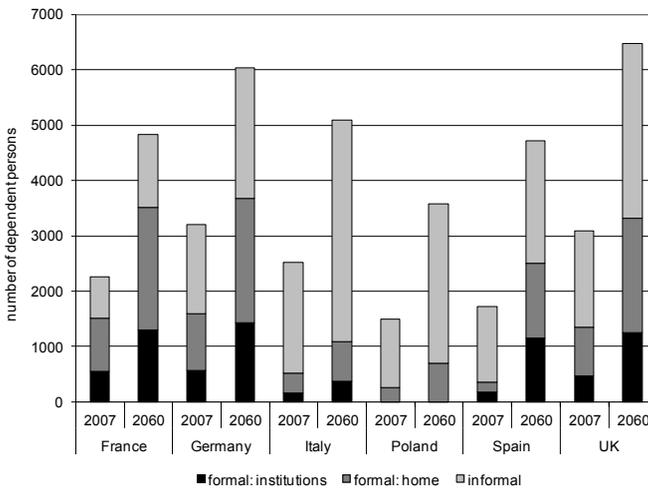
¹¹ In other words, the rate of dependency of an 80-year old in the future is the same as that of an 80-year old today, but there will be more people living up to their 80th birthday in the future than today.

¹² Data source: EC (2009, p. 138).

¹³ Data sources: EC (2009, p. 140).

formal care at home, the percentage change would be 151 (171 for EU10). Finally, the percentage change for those only relying on informal care would be 84 in EU27 (119 in EU10). However, those aggregate trends actually hide strong international heterogeneity, as shown on Figure 2. Whereas the dependent population in France is more or less equally distributed across the three types of LTC, this is not true in many countries. For instance, overall formal care (institutions or home) is nowadays largely underdeveloped in Italy and Poland.

Figure 2. Formal and informal LTC, 2007-2060



Source: EC (2009, p. 138).

We can now turn to the projected public expenditure on LTC presented by the EC *2009 Ageing Report* (EC 2009). Assuming the “pure demographic” scenario, LTC public expenditure is projected to on average increase by 115 percent for EU27. The projected increase ranges from 65 percent in France and the UK to 175 percent and above in the Czech Republic, Spain, Malta, Poland, Romania and Slovakia.

Extrapolating on the basis of existing policies and expenditures does not capture the full scale of the policy challenge. Future changes in the number of people receiving informal or no care, and whether they will receive the care services they need, are also crucial policy questions. Countries with low levels of formal care provision today (and thus low levels of public expenditure), such as Italy, will also witness a very large

increase in the projected number of individuals in need of care. Pressure is likely to emerge in the future for policy changes to increase formal care provision, especially as the future availability of informal care is likely to diminish rather than increase. The gap between the need for care and the supply of formal care will widen due to the growing numbers of elderly people and a likely reduction in the supply of informal care within households.¹⁴

2. The LTC insurance market puzzle

For a large majority of individuals, the cost of LTC in case of severe dependence is high, if not prohibitive. To illustrate this, let us focus on the example of France. Whereas the average pension of a French household is about EUR 1 200 a month, the cost of a good nursing home runs much above that figure. The average cost of institutional long-term care for old people in France is currently at EUR 35 000 per dependent per year (see OECD, 2006), whereas the yearly price of a nursing home in the US ranges between USD 40 000 and USD 75 000 (see Taleyson, 2003).

But despite those large costs, and despite the significant probabilities of becoming dependent in old age (Kemper and Murtaugh, 1991), the LTC private insurance market remains small. This underdevelopment of the LTC private insurance market goes against simple theoretical predictions and, as such, constitutes what is now commonly called the LTC insurance puzzle. In this section, we survey some major factors explaining that puzzle.¹⁵ We shall distinguish between, on the one hand, factors based on perfect rationality and, on the other hand, factors presupposing behavioral imperfections.

2.1 *Excessive costs*

Maybe the simplest explanation of the LTC insurance puzzle lies in its high price which, following standard microeconomics textbooks, leads to a low demand. Various empirical studies show that, for most individuals,

¹⁴ Although the scale of this effect will depend on the starting employment rates of women, among other factors.

¹⁵ Other recent surveys on that issue include Brown and Finkelstein (2011) and Pestieau and Ponthiere (2011).

a private LTC insurance is something very expensive, which they cannot afford. According to Rivlin et al. (1988), only 20 percent of US citizens could purchase insurance for less than 5 percent of their income. Those figures are confirmed by those in the *Lifespans Surveys* (1992): 91 percent of those who do not purchase a private LTC private insurance find that this is far too expensive (see Cutler, 1993). Hence, in the light of this, it appears that explaining the LTC insurance puzzle *a priori* looks quite simple: the high price suffices to explain everything.

The price of a private LTC insurance has been studied in detail by Brown and Finkelstein (2007). Those authors estimate, on the basis of HRS (Health and Retirement Survey) data, that a private LTC insurance purchased at age 65 has a *load factor* equal to 0.18.¹⁶ This means that, for any dollar spent on LTC costs, one can hope to get 0.82 dollars back. That load factor of 0.18 is much larger than for standard health care insurance, where the load factor lies between 0.06 and 0.10 (see Newhouse, 2002). At a first glance, it is tempting to conclude, from those figures, that the low covering rate of the private LTC insurance does no longer look like a puzzle.

However, things are not so simple. Brown and Finkelstein (2007) also provide estimates of the load factor by gender, and find significant differences across gender: men face a load factor of 0.44, whereas women benefit from a load factor equal to -0.04 (that is, better than actuarially fair prices).¹⁷

Quite surprisingly, the participation rate is almost *equal* for men and women, despite that large differential in load factors. That result either suggests that there is a strong correlation, within households, about insurance decisions, or, alternatively, that prices, although high, may not explain the entire picture of the LTC insurance puzzle.

What can explain those high load factors? According to Brown and Finkelstein (2007), it is hard to discriminate, on the basis of empirical observations, between four causes: administrative costs, imperfect competition, asymmetric information and aggregate risk of rising costs. All those causes imply a high loading factor, as well as limits in the benefits

¹⁶ The load factor is defined as the ratio: one minus the expected discounted present value of monetary benefits divided by the expected discounted present value of insurance premia.

¹⁷ This large men/women differential may reflect gender differences in the LTC utilization, due to women's higher longevity but, also, that elderly men are more likely to receive informal aid from their spouses, as compared to elderly women.

comprehensiveness (i.e. quantity rationing), which are also observed (i.e. the typically purchased policy covers only 1/3 of the expected LTC expenditures).

But beyond the actual level of the LTC insurance price, the presence of asymmetric information may make the LTC insurance price even more excessive. Despite recent medical advances, it remains very difficult for an insurer to forecast, for a given individual, the evolution of his autonomy and health status across the lifecycle. In other words, the probability of old age dependency remains very hard to extract for the insurer. At the end of the day, the best informed individual remains the future elderly himself. But if the elderly is more informed than the insurer, the standard *adverse selection* problem arises: only individuals with a sufficiently large probability of old-age dependency will purchase LTC private insurance. Note that this better information about future autonomy prospects may lead individuals to postpone, as much as possible, the purchase of LTC insurance, as suggested by Meier (1999).¹⁸

The existence of asymmetric information on the LTC insurance market is studied by Finkelstein and McGarry (2006). Using AHEAD data (US), they show that there exists no positive correlation between, on the one hand, the fact of having purchased a private LTC insurance and, on the other hand, the probability of institutionalization (nursing home). In other words, the ratio of LTC insured persons to the whole population among individuals admitted to nursing homes does not significantly differ from unity. However, this absence of correlation does not imply that no asymmetric information exists. Finkelstein and McGarry show that the subjective probability of institutionalization within 5 years (which is not observable) is positively correlated with the fact of being insured against LTC. Hence, there exists some asymmetric information, and the absence of correlation between insurance and institutionalization comes from *another* selection mechanism, preferences-based selection, which counterbalances the standard risk-selection mechanism. According to Finkelstein and McGarry, precautionary behavior, as measured by the past purchase of other insurance, is positively correlated with the purchase of LTC insurance, but negatively correlated with institutionalization. This

¹⁸ This alternative explanation of the LTC insurance puzzle is even more likely once it is acknowledged that LTC insurance is associated with large yearly administrative costs (see below). Those costs discourage an early purchase of LTC insurance, and favor postponement.

double selection mechanism explains the contraction of the LTC private insurance market despite the absence of correlation between the purchase of insurance and the institutionalization.

The existence of an adverse selection problem on the LTC insurance market is confirmed by Sloan and Norton (1997). On the basis of two surveys for the US (AHEAD – *Asset and Health Dynamics* – and HRS – *Health and Retirement Survey*), Sloan and Norton find a positive and statistically significant correlation between the subjective probability of entering a nursing home and the probability of purchasing LTC insurance.¹⁹ More recently, Courbage and Roudaut (2008) find, on the basis of SHARE data for France (*Survey on Health, Aging and Retirement in Europe*), that there exists a positive and statistically significant correlation between, on the one hand, having a high risk of dependency (e.g. high BMI scores and high alcohol consumption) and, on the other hand, the purchase of LTC insurance. Hence, the plausible presence of adverse selection may contribute to explain the high LTC insurance costs and, as a consequence, the LTC insurance puzzle.

Note also that, besides that standard, static adverse selection, some studies also highlighted the possible existence of a *dynamic* adverse selection, which would also explain the underdevelopment of the LTC private insurance market. The underlying idea, studied in Finkelstein et al. (2005), is the following. Even in the absence of asymmetric information, there may be an inefficiency of the LTC insurance market in a dynamic environment where agents are strongly encouraged to break their insurance contract. According to the HRS database, the probability of institutionalization is lower among those who break their insurance contract. Note that such a lower probability of institutionalization could, at first glance, be due to standard moral hazard (those keeping the insurance would spend more on LTC).²⁰ However, Finkelstein et al. (2005) reject that hypothesis, on the grounds that, among those leaving the LTC insurance market, the probability of institutionalization is not influenced by the fact of shifting to another insurance, or giving up all insurance. Note, however, that, in the light of the characteristics of those breaking their

¹⁹ Naturally, one cannot exclude the existence of moral hazard explaining that correlation. However, Sloan and Norton (1997) find that family structure variables (marital status and children), which should here affect the occurrence of moral hazard, do not influence the probability of purchasing LTC insurance.

²⁰ On moral hazard in nursing home uses, see Grabowski and Gruber (2007).

insurance contract, explanations alternative to the revelation of new pieces of information may rationalize the break: either initial mistakes by LTC insurance purchasers or, possibly, (uninsured) negative income shocks.

2.2 Unattractive rule of reimbursement (lump sum)

Another explanation for the LTC insurance puzzle consists of the specific *form* of the LTC insurance contracts currently existing on the market. One would expect, like in standard health care insurance, LTC insurance contracts giving a right to the reimbursement of care and service costs, possibly up to a certain limit and with multiple options, including deductibles. The problem is that an increasing number of insurance markets, typically the French one, provide for the payment of a monthly lump-sum cash benefit, which is proportionate to the degree of dependency involved and adjusted according to the evolution of this dependency. These products are closely related to annuitized products and the limited insurance they provide is justified by some type of *ex post* moral hazard.²¹

The degree of dependency can be assessed quite objectively. However, the extent of the needs of a dependent person is much more difficult to assess objectively. The needs of a dependent person are less easy to assess than in the case of well-known disabilities. The perception of LTC is a very recent phenomenon, and the needs implied by a loss of autonomy are vague and susceptible to various interpretations, depending on family background and the social environment. For example, washing difficulties may give rise to a demand for various kinds of services, depending on the person concerned. This diversity of potential needs and demands may give rise to a large number of costly discussions. Hence, to avoid these discussions, insurance firms prefer to offer a cash benefit that dependent people can use in their own way, with the consequence that some individuals feel shorthanded.

Although the precise form of LTC insurance contract may seem irrelevant for the explanation of the LTC insurance puzzle, the associated incompleteness of the LTC insurance contract has often be proposed as a major cause of the underdevelopment of private LTC insurance markets.

²¹ This type of moral hazard cannot allegedly be taken care of by the traditional co-payments or deductibles.

In a pioneer paper, Cutler (1993) argued that, since there exists a long delay between, on the one hand, the purchase of the LTC insurance and, on the other hand, the first LTC-related costs, there may exist a strong intertemporal variability of LTC costs per dependent person. The risk of a rise in LTC costs per dependent person is common to all members of a cohort and thus, cannot be diversified on a cohort (contrary to the risk of loss of autonomy, which can be diversified on a single cohort). Therefore, the only way for insurers of protecting themselves against too large reimbursements to elderly dependent due to a rise in LTC costs is to carry out intertemporal pooling on several cohorts. But such an intertemporal pooling can only work when LTC costs are not intertemporally correlated. Unfortunately, LTC costs are also strongly correlated over time, which makes intertemporal pooling difficult. As a consequence, the risk of a rise in LTC costs per dependent cannot be fully insured. This explains why LTC insurance contracts now propose lump-sum reimbursement, or numerous limitations to reimbursement (e.g. thresholds). Moreover, Cutler (1993) also emphasizes the strongly risky nature of LTC insurance. This highly risky nature has an immediate corollary for investors: large risk premiums are required, to compensate for the risk taken by the insurers. Those high risk premiums lead to excessive LTC insurance prices (see below).

In sum, the inadequacy of lump-sum reimbursement deters individuals from purchasing a private LTC insurance.²² That explanation differs from the previous one, which presupposed a complete LTC insurance, but questioned the high level of the insurance price. On the contrary, the core of that alternative explanation of the LTC insurance puzzle lies in the necessarily *incomplete* nature of the LTC insurance contract. Rational forward-looking individuals may not want to purchase such an incomplete insurance.²³

²² However, some studies, such as Taleyson (2003) and Kessler (2007), argued, contrary to Cutler (1993), that lump-sum reimbursement is a major factor explaining the dynamism of the French LTC insurance market as compared to the US market.

²³ The formal properties of the fixed reimbursement insurance are studied in detail in Eeckhoudt et al. (2003). See also Cremer, Lozachmeur and Pestieau (2012)

2.3 Crowding out by the family

Another explanation of the LTC insurance puzzle consists of crowding out of private LTC insurance by family solidarity.²⁴ The underlying rationale goes as follows. True, if economic agents were living alone, without any family or friends, there would exist few effective ways of insuring oneself against the substantial – and highly likely – LTC spending at old age. Therefore, in such a narrow context, forward-looking rational agents would definitely buy private LTC insurance. However, the real world is quite different, and many individuals can, once dependent, rely on their spouse or on their children to be helped in case of LTC. Therefore, the low level of private LTC insurance coverage does not result from irrationality. On the contrary, private LTC insurance is regarded as non-optimal by rational individuals who anticipate future help from their family.

Another family-based explanation of the LTC insurance puzzle was proposed by Pauly (1990). Parents actually prefer not to be sent to an institution once they are dependent. Clearly, parents have a strong preference for receiving help from their own children or grandchildren. That preference for family-provided LTC tends, under weak assumptions, to rule out LTC private insurance. Indeed, such insurance tends to reduce the cost of institutionalization and, hence, increases the probability of being sent to a nursing home. Therefore, provided that, in case of old-age dependency, a rational person prefers to be helped by a family member rather than by an unknown social worker, the incentive to purchase a private LTC insurance is low, even in the absence of state assistance.

It should be stressed, however, that the existence of family concerns does not, on its own, suffice to explain the LTC insurance puzzle. The reason is that the precise form of parental *preferences* matters. If a parent is sufficiently altruistic towards his children, then, as argued by Pauly (1996), he will buy LTC private insurance in order to avoid burdening his spouse or children in case of old-age dependency.²⁵ He will do so despite the fact that, from a purely egoistic perspective, he would have preferred being helped by his children or spouse rather than being sent to an anon-

²⁴ That explanation is quite close to the one used in another well-known puzzle: the annuity puzzle (see Brown, 2007). In that context, the family would provide an insurance against an unexpected long life.

²⁵ This burden could be in terms of money (LTC spending) or time (in case of home-based informal help).

ymous institution. However, if a parent is not sufficiently altruistic, he will behave strategically and use the promise of high bequests to be helped by his family (see Norton, 2000).²⁶ Therefore, the mere existence of family concerns only explains the LTC insurance puzzle under particular preferences for parents.

This family-based explanation of the LTC insurance puzzle has been subject to various empirical tests, with quite equivocal results. On the basis of US data, Sloan and Norton (1997) show that the bequests left to descendants do not have any statistically significant effect on the demand for private LTC insurance. That empirical finding does not support the family-based explanation. However, a more recent study by Courbage and Roudaut (2008) shows, on the basis of the French SHARE data, that being married and having children makes it more likely to purchase private LTC insurance. This latter empirical result supports the importance of parental preferences for the issue at stake (Pauly, 1996), but without fully validating the family-based explanation of the LTC insurance puzzle. Indeed, if parental altruism makes parents buy LTC private insurance, then the underdevelopment of LTC insurance would reveal the widespread lack of altruism among parents, which does not seem fully convincing.

2.4 Crowding out by the state

Besides the reliance on the family, another possible explanation for the LTC insurance puzzle points to a potential crowding out by state assistance (Norton, 2000). The underlying idea is the following. By acting as the Good Samaritan, the government can supply some aid to the dependent elderly without resources. Therefore, rational forward-looking individuals have little incentive to buy LTC private insurance, simply because they can benefit from state-provided resources at old age without having purchased any private insurance. Hence, provided that the state can help the elderly dependent in need, buying a private LTC insurance is a waste of resources.

Note that this kind of crowding out argument does not require the actual existence of a large public LTC program. Indeed, only the *expecta-*

²⁶ We assume that children cannot buy an insurance to protect themselves against LTC spending on their parents.

tion of state-provided help to the elderly dependent in need suffices to do the job. Hence, given the distance to old-age dependency, it is easy for individuals to believe that, by the time they become dependent, the state will have developed a new “pillar” of the social security system.²⁷

The hypothesis of crowding out by the state has been largely debated in the US, where social assistance – i.e. Medicaid – is often suspected to be at the origin of the LTC insurance puzzle. Sloan and Norton (1997) show that there exists a statistically significant negative correlation between, on the one hand, the probability of purchasing a private LTC insurance and, on the other hand, the variables determining the eligibility for Medicaid. More recently, Brown and Finkelstein (2008) estimate a lifecycle utility model for an individual of age 65 (men and women) choosing a lifecycle consumption profile under risk for LTC expenditures. They compute the willingness to pay for LTC private insurance under various degrees of risk-aversion and show that, for a wide range of preferences, the utility gain from buying LTC insurance is negative. They also argue that Medicaid, by its role of last resort payer, would explain at least 2/3 of the contraction of the US private insurance market, even when actuarially fair LTC insurance would be available. The hypothesis of crowding out by Medicaid has also been tested by Brown et al. (2007). On the basis of HRS data, they estimate that a USD 10 000 fall of the Medicaid eligibility threshold would increase the LTC insurance coverage ratio by 1.1 points.²⁸

In the light of those results, the crowding out of private LTC insurance by Medicaid in the US seems to be statistically significant. However, the exact size of the crowding out phenomenon remains hard to quantify. This leaves us with two possibilities. Either the crowding out by the state assistance only explains a part of the LTC insurance puzzle; i.e. other factors are also at work. Alternatively, it may be that, rather than the actual assistance by the state, the crowding out may follow from expectations about future state assistance.

To conclude, it should be stressed that the crowding out by state assistance, if it exists, can take various forms, which have different implica-

²⁷ See Section 4 on the difficulties raised by the construction of that pillar.

²⁸ According to Brown et al. (2007), the minor effect of eligibility criteria can be explained as follows. Provided that Medicaid remains a secondary payer, it follows that even without any asset limits to Medicaid eligibility, a large portion of the private insurance benefits are redundant to what Medicaid would otherwise have paid.

tions. First, some individuals may decide to spend all their wealth when being healthy, in order to become eligible for state assistance. Second, one may simply hide one's resources in such a way as to become eligible for social assistance. Third, some individuals may, in contrast, transfer their wealth to their children through *inter vivos* gifts, in order, here again, to become eligible for social assistance. Those three cases are characterized by different individual lifecycle consumption profiles, for individuals and their siblings. But, from the point of view of the state, the outcome is the same: such a strategic behavior makes those individuals eligible for assistance, through means-tested benefits such as Medicaid in the US or the APA in France. Given that the public authorities are often reluctant to reclaim part of the estate of those having benefited from LTC assistance, such strategic behaviors will be adopted and lead to the effective crowding out of private LTC insurance by the state.

2.5 State dependent utility

So far, we have implicitly assumed that the form of the utility function is the same in all states of the world. LTC prevalence is an instance when this assumption might be violated. The preferences are totally different for someone who is healthy and has a variety of goals in life, and for someone who is disabled, and has well-defined but limited needs. In standard utility maximization problems, the first-order conditions that characterize the optimum equate the marginal utility of consumption across states. If the utility function is the same in all states, an individual equates the marginal utility of consumption by equating the level of consumption across states. Insurance provides a simple mechanism for smoothing consumption across states of the world. Assume that to reach a certain level of welfare, one needs much more resources in a state of disability than in a state of full autonomy and that in case of disability, one quickly reaches some type of satiation (there is a limit to the level of nursing one's needs). In that case, the demand for insurance can be nil or, at best, low.

Let us denote the utility in the state of autonomy as $u(c)$ and the utility in the state of dependence as $H(m)$, where c and m represent the corresponding level of consumption. The probability of dependence is π and θ

is the premium for an actuarially fair insurance. The expected utility can be written as

$$EU = u(w - \theta)\pi + H(w - \theta + \theta/\pi)(1 - \pi) \quad (1)$$

where w is the initial wealth. Clearly, it pays to buy some insurance if $-u'(w) + H'(w) > 0$. In other words, if the marginal utility of wealth is higher in disability than in autonomy. This in turn may depend on the wealth level. When H reaches a saturation level, $u'(w)$ may be larger than $H'(w)$ when the individual is sufficiently wealthy. In other words, once nursing is financed, the wealthy disabled has no more need.

Finkelstein et al. (2008) note that it is *a priori* ambiguous whether the marginal utility of consumption rises or falls with deteriorating health, given that some goods (e.g. travel) are complements to good health while other goods (e.g. assistance with self-care) are substitutes for good health. However, they also provide evidence, using subjective well-being measures from HRS, that a one standard-deviation increase in an individual's number of chronic diseases is associated with an 11 percent decline in marginal utility. They report that this reduces the optimal share of medical expenditures covered by health insurance by about 20 to 45 percentage points. In a theoretical paper, Bien et al. (2012) derive the conditions under which rational forward-looking agents do not buy LTC private insurance; these conditions pertain to the substitutability between three dimensions of welfare: consumption, autonomy and health.

2.6 Myopia or ignorance

So far, we have only considered explanations of the LTC insurance puzzle, which suppose *rational forward-looking* agents. Put differently, when deciding not to purchase private LTC insurance, individuals do not make any mistake or judgment error. According to those explanations, it is rational, in the presence of a high LTC insurance loading factor, not to buy an insurance that only provides quite limited reimbursement under the form of lump-sum payments. The presence of family and state assistance reinforces the incentive not to purchase private insurance. Finally, if state-dependent utility is such that, under old-age dependency, the marginal utility of income is very low, there is also no rational argument for

purchasing private LTC insurance. Hence, under those explanations, agents behave rationally and, as a consequence, the low level of LTC private insurance markets is also individually rational.

Whereas those explanations are plausible, these are not the unique possible ones. It is also possible to rationalize the observed low covering of LTC insurance, while regarding the non-purchasing decisions as *irrational*. This kind of explanation can take various forms, but each of these involves some kind of behavioral imperfection.

A first behavioral explanation consists of an *underestimation*, among the population, of the risk of old-age dependency. There is a well-known downward bias of the probability of occurrence of negative events in life. Old-age dependency obviously being negatively loaded, individuals are likely to minimize its frequency of occurrence. Note, however, that such a downward bias is not benign at all as far as the demand for LTC private insurance is concerned. Under a low probability of old-age dependency, the individual's incentive to purchase insurance is pretty low since the expected welfare gains from such an insurance are not only temporally distant, but, also, highly unlikely, whereas the cost of such an insurance is certain (and high, as shown above). Thus, some underestimation of the risk of old-age dependency may explain a significant part of the LTC insurance puzzle.

The objective probabilities of old-age dependency estimated in the literature are quite high. For instance, according to Murtaugh et al. (1997), an individual aged 65 has a 0.43 probability of entering a nursing home. That probability is also shown to differ significantly across gender: it is equal to 0.33 for men (as their wife will generally be in better health and thus will take care of them), and to about 0.50 for women. Moreover, Murtaugh et al. (1997) show that the stays at nursing homes are long: 15-20 percent of the newcomers will remain more than five years. Taken together, those estimates should, in principle, make a large proportion of the population at risk buy LTC private insurance. Naturally, that claim presupposes that individuals are well-informed and can easily manipulate probabilities.

On the basis of the high objective probabilities of old-age dependency, one can interpret the low demand for LTC insurance as revealing the downward bias in the subjective probabilities of old-age dependency. Finkelstein and McGarry (2006) show, on the basis of AHEAD data (av-

erage age: 79 years), that the distribution of the subjective probability of entering a nursing home within the next five years of life has a singular form, and is not single-peaked. About 50 percent of the population consider that the probability that they will enter a nursing home in the next five years is zero. The second peak of the distribution arises at the value of 0.50: about 15 percent of the population believe that the probability of entering a nursing home equals 0.50. Very few people assign a probability larger than 0.50. Undoubtedly, that singular distribution of the subjective probability of old-age dependency supports the underestimation thesis: the low demand for LTC insurance would thus reveal individuals' – excessive – optimism about future health status.

It should be stressed, however, that subjective probabilities may not explain the LTC insurance puzzle as a whole. To see this, note that, according to Finkelstein and McGarry (2006), eight percent of the individuals who believe that they will definitely not enter a nursing home within 5 years have actually purchased LTC insurance. Given the high age of the surveyed individuals, this figure is somewhat surprising, and suggests that individuals may have difficulties in manipulating small numbers, and in drawing all conclusions from their subjective beliefs.

More importantly, biases in the assessment of old-age dependency risk may tend to vanish over time, as individuals can, over their lifecycle, learn about their health capital, for instance by observing the health of their own, elderly parents. Such a *learning* process may tend to qualify the underestimation hypothesis, by suggesting that this cause of the LTC insurance puzzle would only be valid in a very short time horizon. Regarding this learning effect, Courbage and Roudaut (2008) report, on the basis of French data in SHARE, that the probability of purchasing a private LTC insurance is increasing with the fact of having received an informal help, and is also increasing in the fact of having provided such a help. Those empirical findings cast some light on the formation of subjective beliefs about old-age dependency risk.

In sum, the existing literature – the observed gap between objective and subjective probabilities of old-age dependency – suggests that there may be a strong behavioral explanation to the LTC insurance puzzle. Nonetheless, the precise form of the behavioral imperfection is harder to identify. Low subjective probabilities of old-age dependency may reflect myopia, or ignorance, or optimism, or some other bizarre attitude to risk.

Having stressed this, one can hardly, in the light of the existing empirical literature, keep the standard objective expected utility models with full information as good approximations for the description of real choices in terms of LTC insurance.

2.7 Denial of heavy dependence

Finally, besides the subjective *versus* objective probability issue, another behavioral explanation can be explored here. Clearly, when discussing LTC so far, we did *as if* old-age dependency is regarded as a standard, everyday life issue by individual decision-makers. In that context, purchasing a private LTC insurance would be formally close to purchasing common goods (like cars etc.). This constitutes an obvious simplification. Old-age dependency is the exact opposite of an everyday life issue. Old-age dependency is a unique event in one's life (i.e. something comparable to childhood). Given the singularity of old-age dependency, one can hardly treat the purchase of a private insurance against LTC costs as the purchase of a normal insurance (e.g. against car accidents or domestic fires).

Heavy dependence, like death, is a source of anxiety and, as such, this may imply the possibility of denial of dependence-relevant information, interacting with intertemporal choices. One would try to forget about old-age dependency in the same way as one tries to forget about death. Such a denial of old-age dependency is likely to lead to time-inconsistent decisions and other "behavioral" phenomena.²⁹

The repression of signals of mortality leads to underinsurance for unsophisticated individuals. Note that for forward-sophisticated individuals, the result can be reversed: they may over-insure in anticipation of future denial and seek commitment devices. The refusal to face up to the reality of dependence may help explain an inadequate purchase of LTC insurance. Private LTC insurance only makes sense provided that one acknowledges the mere existence of old-age dependency. Denying that possible event in life makes the purchase of LTC insurance so irrelevant that it will not even enter the set of possible consumption bundles.

Although that denial explanation of the LTC insurance puzzle shares some psychological, behavioral features with myopia or ignorance (see above), one should be careful to avoid mixing those two types of explana-

²⁹ On the denial of death and its behavioral consequences, see Kopczuk and Slemrod (2005).

tions. Clearly, while one may think about (more or less) easy ways of correcting for myopia or ignorance (e.g. information campaigns, adequate taxes or subsidies), the same is not that obvious in case of denial of old-age dependency. The reason is that a denial is not due to the failure to perceive things, or to collect or process the required information. On the contrary, a denial consists of a lack of will to do so. This kind of behavioral imperfection seems harder to overcome by means of standard policy instruments.

3. The role of family solidarity

As already mentioned, most seniors with impairments reside in their home or that of their relatives, and they largely rely on volunteer care from the family. These include seniors with severe impairments (unable to perform at least four activities of daily living). And many people who pay for care in their home also rely on some donated services. The economic value of volunteer care is significant, although the estimates of it are highly uncertain.

Whether this solidarity is sustainable at its current level is an important question. There are numerous sources of concern. The drastic change in family values, the increasing number of childless households, the mobility of children, the increasing labor participation of women are as many factors explaining why the number of dependent elderly who cannot count on family solidarity is increasing.³⁰

An important feature that is often neglected is the real *motivation* for family solidarity. For long, we have adopted the fairytale view of children or spouses helping their dependent parents with joy and dedication, what we call pure altruism. We now increasingly realize that family solidarity is often based on forced altruism (social norm) or strategic considerations (reciprocal altruism). In this section, we review some recent work on these issues.

Knowing the foundation of altruism is very important in order to see how family assistance will react to the emergence of a private or public scheme of LTC insurance. For example, the introduction of LTC social

³⁰ See Duée et al. (2005) on the factors explaining the rise in the number of elderly dependent individuals in need who cannot benefit from someone's informal help.

insurance is expected to crowd out family solidarity based on pure altruism, but not necessarily that based on forced altruism. In families where solidarity is based on strategic exchanges (bequest or *inter vivos* gifts in exchange for assistance), the incidence of a social LTC scheme will be a decline in intergenerational transfers. The issue of crowding out is pervasive as it does not only concern the possible substitutability between family solidarity and formal schemes, but also between social and private LTC, as we will see below. We now survey a sample of recent microeconomic papers dealing with LTC and family relations.

3.1 Strategic bequest motive

The classic paper on strategic bequests by Bernheim et al. (1985) shows that parents can, in theory, extract from their children the maximum amount of attention and/or assistance, by playing them against each other with the prospect of inheritance. In this type of model, parents have a hold on the game. That strategic bequest motive is shown, in that same paper, to be empirically supported by the existence of a positive correlation between the attention paid to parents (the number of visits and phone calls) and the (potential) wealth inherited by children.

Note, however, that the strategic bequest motive presupposes that the dependent elderly has sufficiently good cognitive skills. The reason is that the elderly dependent can only take part in exchanges with his children provided that he remains the effective decision-maker regarding the allocation of his resources. Thus, under the strategic bequest motive, children's care is conditional on parental cognitive awareness, unlike under pure family altruism.³¹

This observation has been used in subsequent empirical studies in order to test the existence of a strategic bequest motive. Hoerger et al. (1996) show, on the basis of *National Long Term Care Survey* (NLTCS) data, that the level of parental cognitive awareness has no impact at all on the attention and care received from children. It even appears that, among cognitively able parents, the received informal help is smaller among

³¹ For cognitive skills along the lifecycle, see recent studies by Adam et al. (2007) and Agarwal et al. (2007). The cognitive performance decreases beyond the age of 60, with slopes varying with the socio-professional group.

richer parents. This contradicts the existence of a strategic bequest motive.³²

Sloan et al. (1997) consider an alternative family model where the elderly parent derives some welfare from the child's informal aid. The child is not purely altruistic, but he would like his parent to be helped. The levels of informal and formal care are then treated as the outcome of an intrafamily bargaining process. Informal aid is, in theory, increasing with the number of children (as this makes the threat of no bequest more plausible), increasing in the degree of parental cognitive awareness (see above), and in the parent's wealth (since this makes the child follow his parent's will better). Empirical tests using NLTCs data lead to the rejection of all those theoretical predictions. Thus, there is little empirical support for the hypothesis that care giving by children is motivated by the prospect of receiving bequests from their parents. Moreover, Sloan et al. (1997) show that Medicaid subsidies have not "crowded out" informal care provided by relatives and friends of the dependent elderly, nor have they reduced wealth accumulation by the elderly.

3.2 Family games: The dependent elderly and his children

A central aspect of LTC consists of the precise living arrangement concerning the dependent elderly. Hoerger et al. (1996) develop a model where the family chooses a living arrangement, in such a way as to maximize family welfare subject to the family budget constraint. Three living arrangements are considered: (1) the dependent elderly lives at home; (2) the dependent elderly lives in a nursing home; and (3) the dependent elderly lives in an intergenerational household. Hoerger et al. (1996) also examine the effects of public subsidies on the living arrangements of the disabled elderly. Direct subsidies for nursing home care and state policies which significantly limit nursing home beds or reimbursement affect the choice of living arrangement. State policies which subsidize community living have little effect on nursing home entry, although they increase the probability of living independently.

Hiedemann and Stern (1999)³³ argued that, in the previous literature on living arrangements for the elderly, the number of children was largely

³² See also Sloan et al. (1997) for similar results.

³³ See more recently Hiedemann et al. (2012).

ignored. As a consequence, little attention was paid to strategic interactions among children. In order to study those interactions among children, Hiedemann and Stern develop a family model with one dependent elderly and several children. Three living arrangements are available: (1) the dependent parent lives alone; (2) only one child (the primary caregiver) provides some care to his parent; (3) the dependent parent is sent to a nursing home.

The decision process has two stages. In the first stage, each child decides whether he proposes to serve as the primary caregiver. Then, in a second stage, the dependent parent chooses the best living arrangement among the available ones (i.e. children's bid in the first stage, plus living alone or in a nursing home). Given that the child proposes his service as a primary caregiver only if there is a net positive expected utility gain from such a proposition, the first-stage decision obviously depends on the parent's preferences (in particular on the probability that he regards the child's proposal as the best).

Assuming, in the first stage, that only one component of the direct utility assigned by a particular child to a particular living arrangement is common knowledge, whereas the second component is a stochastic component, Hiedemann and Stern solve the model numerically, under different structures of preferences (e.g. free riding by children, or, alternatively, children with exclusivity tastes for care giving). They derive the equilibrium, which here consists of a vector of each child's probability of proposing his help, such that all beliefs of all children are consistent with each other. Hiedemann and Stern then estimate, on the basis of NLTC data, the determinants of the family member's utility. They show that dependent women value children's informal help more than dependent men, and that the help of a caregiver of the opposite gender is, on average, more valued by the dependent parent. More importantly, Hiedemann and Stern show that two imperfections prevent the maximization of aggregate welfare: imperfect information and coordination failures.

Stern and Engers (2002) compare two family decision models. On the one hand, the unitary household model, where the entire family chooses the best living arrangement for the dependent elderly, that is, the living arrangement maximizing the family (aggregate) utility function. On the other hand, the voluntary model, where each family member is allowed to choose not to take part in the caring decisions. In that second model, the

chosen option is the one that maximizes the total utility of the participating family members. Stern and Engers also rely on data from the NLTCs to estimate and test the parameters of their models. Then, they use the parameter estimates to simulate the effects of the existing long-term trends in terms of the common but untested explanations for them. They also simulate the effects of alternative family bargaining rules on individual utility to measure the sensitivity of their results to the family decision-making assumptions they adopt.

More recently, Pezzin et al. (2007) use a two-stage bargaining model to analyze the living arrangement of a disabled elderly parent and the assistance provided to the parent by her adult children. In the first stage, the living arrangement of the parent is chosen (living alone, in a nursing home, or with a child). In the second stage, the assistance/transfers from the children to the elderly dependent are determined, given the living arrangement chosen in the first stage.

Working by backward induction, Pezzin et al. (2007) first calculate the level of assistance that each child would provide to the parent in each possible living arrangement. Using these calculations, they then analyze the living arrangement that would emerge from the first-stage game. Various types of games are considered. One major result is that, since coresidence tends to reduce the bargaining power of coresident children (and, thus, to put him in a bad situation in stage 2), that living arrangement, though Pareto-efficient, is unlikely to emerge from the first stage in the absence of explicit commitment about assistance at stage 1. That is, the outcome of the two-stage game need not be Pareto efficient.

Whereas most game-theoretical family models of LTC provision have either focused on the children-parent relationship, or on the strategic interactions among children, an important exception is Pezzin et al. (2009) who study whether the presence of children has any impact on the LTC assistance provided by a parent to his/her dependent spouse.

3.3 Family solidarity in a dynamic world

Finally, whereas the articles surveyed so far study the behavior of family members in a static context, two recent papers focus on the role of the family in the provision of LTC in a dynamic set up. Canta and Pestieau (2012) develop an overlapping generations model where each cohort of

adult children is divided between, on the one hand, traditional agents, who provide an amount of LTC help equal to what was provided by the previous generation and, on the other hand, opportunistic agents who choose LTC by maximizing their expected lifetime welfare, while anticipating that, if they have traditional children, providing low LTC to their parent will imply low LTC received once being dependent. Assuming exogenous transition probabilities across types, Canta and Pestieau then characterize the optimal LTC policy at the stationary equilibrium, and discuss the occurrence of crowding out of private LTC insurance in their results. They find two reasons for public action: redistribution and an informational externality.

In another recent paper, Ponthiere (2011) examines whether state-provided LTC can crowd out family-based LTC in an economy where each adult cohort is divided between altruistic and non-altruistic individuals, and where the transmission of the altruistic trait follows a socialization process à la Bisin and Verdier (2001). Ponthiere shows that if the state provides LTC to dependent parents who are not helped by their children, parental socialization efforts aimed at transmitting altruism are necessarily reduced. However, that fall in the investment in the values of children does not necessarily lead to a crowding out effect. The existence of crowding out depends on the form of the dependent's utility function, on the cost and form of the socialization mechanism at work, and may also depend on the initial conditions (i.e. the current prevalence of family altruism).

In sum, although non exhaustive, this short survey suggests that considerable efforts have recently been made to open the family "black box", in such a way as to represent it as a set of interdependent individuals, whose behavior in terms of LTC provision, dependent living arrangements, or own spatial localization, is not necessarily based on pure altruism but, also, on strategic considerations or social norms.

4. Social LTC insurance

There are very few countries with explicit LTC social insurance programs. Furthermore, these programs are not very generous: they only cover a small fraction of the LTC cost (typically EUR 500 a month) and

yet their sustainability is uncertain. Let us as an example describe the most developed of these schemes, the German one that was introduced in 1995 and has been coined as the “5th pillar” to the social security system.³⁴ This LTC insurance covers the risk of becoming dependent on nursing care and attention and it is taken out with the company providing health insurance. If the individual is covered by state health insurance, he automatically has long-term care insurance. If he has private health insurance and is entitled to general hospital care, he also has private long-term care insurance. As for health insurance, public long-term care insurance is financed through contributions of 1.7 percent of the gross salary split equally between employer and employee. Employers deduct contributions directly from the wages and transfer them to the health insurance funds.

To be fair, in most countries, health care systems cover the medical aspects of dependence and the assistance side of social protection provides means tested LTC nursing services. The best-known example of that is the American Medicaid that is suspected to discourage the development of an efficient market for LTC insurance. As we have seen above, there exists some work on this issue, mostly empirical. There is little theoretical work on the issue of LTC social insurance.

To approach this issue, one has to consider a social planner with some objective function comprising equity and efficiency aspects. This planner typically acts as a Stackelberg leader, that is, it can commit to a policy and anticipates the supply and demand responses of individuals and the behavior of families and private insurers. In other words, both families and markets act as followers. If, by any chance, market forces and family solidarity yield a desirable outcome, our central planner does not intervene. A few questions have already been addressed in this area. Jousten et al. (2005) focus on families with different levels of altruism. Given the cost of public funds, the central planner tries to induce the more altruistic families to assist their dependent parents and only provide aid to the dependent elderly whose children are less altruistic. This may imply a suboptimal quality of public LTC. Pestieau and Sato (2006, 2008) study the problem of evenly altruistic children who differ in their earning capacities, that is in their wages and thus, in the opportunity cost of the time they can volunteer assisting their dependent parents. In case of parents’ dependency, the more productive children will tend to provide financial

³⁴ The first four pillars are: health, family, unemployment and retirement.

help whereas the less productive children will opt for assistance in time. Parents with sufficiently high pensions or other resources and who do not expect enough assistance from their children will purchase some private insurance. The social welfare maximizing government can subsidize either private assistance or private insurance; it can also directly provide nursing services. The final outcome is shown to depend on the loading cost of private insurance, the cost of public funds and the wealth of the parents.³⁵

In the remainder of this section, we present a great deal of ongoing research on social insurance for LTC. But first of all we discuss some conceptual issues underlying the assumptions that have to be made when modeling a public LTC scheme.

4.1 Conceptual issues

In designing an optimal public insurance scheme for LTC, one faces a number of conceptual hurdles. We first need to define the social welfare criterion that will be used in such a design. In particular, if we are concerned by the wellbeing of both parents and children and if children help their parents in case of disability out of altruism, it must be decided whether or not the altruistic component of children's utility has to be taken into account or simply laundered out. Laundering out filial altruism is often advocated to avoid double counting. If the altruistic part of the children's utility were kept, the welfare of the disabled parents would be attributed too much weight.

Another issue arises when the aid coming from children is not motivated by altruism but by a social norm. In that case, family solidarity should not only be laundered out but could even be negatively weighted to account for the indirect cost inflicted on the aiding persons.

One delicate ethical issue is how to treat the state of severe dependence, which is a state where the dependent are unable to recognize those close to them or even their own reflection. This is a state where care is needed but not acknowledged. It might be tempting to think that the value of life in that state is lower than when the dependent are fully conscious.

There is also the issue of (new) paternalism that arises in case of misperception. Typically, if the individuals ignore or underestimate the risk

³⁵ The assistance from children decreases if the parents' wealth increases.

of disability in old age, one expects the government to induce them or even force them to take the appropriate protective steps.

Finally, there is the pitfall of utilitarianism when dealing with different preferences. Consider two individuals who live for two periods. Individual 1 will be autonomous all his life; individual 2 is expected to be disabled in the second period of his life. Denoting the utility of consumption by $u(c)$ for the first period and $u(d)$ for the second period and that of long-term care by $H(m)$, the problem of the social planner is to maximize:

$$u(c_1) + u(c_2) + u(d_1) + H(m_2), \quad (2)$$

subject to the resource constraint: $c_1 + c_2 + d_1 + m_2 = Y$, where Y represents the available resources and both the rate of interest and the time discount rate are 0. From the FOCs, one obtains that $c_1 = c_2 = d_1 = c \geq m_2$ depending on the utility functions.

The question of social insurance for LTC is very complex and it is impossible to deal with all issues at the same time. Consequently, we will divide it into four problems: (i) social insurance with uncertain family solidarity; (ii) social insurance along with private insurance; (iii) social assistance with strategic impoverishment; and (iv) social insurance determined by majority voting. First, we present what can be viewed as a canonical model.

Consider the lifetime utility of an individual facing a double uncertainty: that of dependence, π , and (conditional on being dependent) that of getting his child's aid, p . In the first period, the individual consumes c and enjoys a utility level of $u(c)$. He devotes e units of time to boost the probability of altruism of his child and the rest, $1 - e$, to market labor at wage w . His earnings $w(1 - e)$ finance his consumption c , saving s , and private insurance θ . He faces a probability π of becoming dependent. Thus, with probability $1 - \pi$, he will have a healthy second period and consume d , which comes from the proceeds of his saving, Rs . His utility is given by $u(d)$. In case of dependence, his welfare is given by $H(m)$ where m denotes the amount of LTC. One expects that $u(x) > H(x)$, namely to reach the same level of utility, one needs more resources in a state of disability than in a state of autonomy. In case of dependence, the individual will receive some help from his child with probability $p(e)$.

In that case, the level of care he receives is given by $m_1 = Rs + g + a + \theta\lambda/\pi$, where a is the level of help g , the social benefit financed by a payroll tax of rate τ , and $\theta\lambda/\pi$ the compensation from private insurance with λ being the loading factor. The level of aid a results from an altruistic choice by the child. With probability $1 - p$, the parent does not get any aid from his child and his level of care is $m_0 = Rs + g + \theta\lambda/\pi$. We can now write the lifetime utility of the parent:

$$U = u(c) + (1 - \pi)u(d) + \pi p H(m_1) + \pi(1 - p)H(m_0), \quad (3)$$

or

$$U = u(w(1 - \tau)(1 - e) - s - \theta) + (1 - \pi)u(Rs) + \pi p(e)H(Rs + g + a + \theta\lambda/\pi) + \pi(1 - p(e))H(Rs + g + \theta\lambda/\pi). \quad (4)$$

The revenue constraint implies that

$$w\tau(1 - e) = \pi g. \quad (5)$$

We now look at four different models. In the first model, private insurance is assumed away. The problem is the design of a LTC scheme (τ, g) that maximizes the lifetime utility of the individual. Heterogeneity of w and p is allowed. The second model focuses on the relation between private and public insurance. Now, family solidarity is assumed away and the main source of heterogeneity is w . Misperception is introduced, in which case the degree of misperception is another source of heterogeneity. The third model deals with the problem of relatively wealthy individuals who could afford to finance their LTC but prefer to benefit from means-tested schemes to increase their bequest capacity. This is done through what is called a strategic impoverishment. Whereas the first three models are normative, the fourth is positive. Society is made of individuals who differ in w , p and π and we look at the existence of majority voting on g .

4.2 Four problems

Public long-term care scheme with uncertain altruism

Cremer, Gahvari and Pestieau (2012) consider a society with two overlapping generations. There are two periods. In the first the parents work, educate their children and save for retirement. In the second, parents retire and incur a certain risk of dependence. Their children work and provide them with assistance in case of loss of autonomy. For various reasons pertaining to migration, health and goodwill, there can here be a default of assistance. The probability of assistance is assumed to partially depend on the time that parents have earlier devoted to their children. The LTC insurance market is assumed away but there is a possibility of developing a social insurance program for LTC. This means that as a protection against dependence, the parents can count on both their children's assistance and social benefits. They can also count on their own saving; that is, they can self-insure.

The authors look at the case of a representative individual and consider a three-stage game. The first stage is the choice by the government of social benefit and a tax to finance it. The second stage is the choice by the parent of education and saving. The third stage is the choice by children of the level of assistance the dependent parent gets. Children's assistance is uncertain; it has a probability that depends on the time that parents devote to shaping their preferences and inculcating them with values. In the *laissez-faire*, LTC will be underprovided; its level will be insufficient for individuals who do not receive assistance from their children. There is thus room for government intervention even with (*ex ante*) identical individuals. The main result is a tax formula which comprises an insurance term and an efficiency term. The tax will depend positively on the gap between the marginal utility of LTC in case of non-assistance and the marginal utility of consumption in the first period and negatively on the compensated effect of the tax on the time devoted to one's children.

Long-term care private and public insurance

One of the main rationales for social insurance is redistribution. Starting with the paper by Rochet (1991), the intuition is the following. We have an actuarially fair private insurance and the possibility of a social insurance scheme to be developed along with an income tax. If there were no

tax distortion, the optimal policy is to redistribute income through income taxation and let individuals purchase the private insurance that fits their needs. If there is a tax distortion, and if the probability of loss is inversely correlated with earnings, then social insurance becomes desirable. Given that low-income individuals will benefit more from social insurance in a distortionless way than high-income individuals, social insurance dominates income taxation. In that reasoning, moral hazard is assumed away but the argument remains valid with some moral hazard.

While the above proposition applies to a number of lifecycle risks, it does not apply to risks where the probability is positively correlated with earnings. This seems to be the case for LTC.³⁶ Dependence is known to increase with longevity and longevity with income. Consequently, the need for LTC is likely to be positively correlated with income, and Rochet's argument implies that a LTC social insurance would not be desirable. This statement does not seem to fit reality, where we see the needs for LTC at the bottom of the income distribution. Where is the problem?

First, we do not live in a world where income taxation is optimal. Second, even if we had an optimal tax policy, it is not clear that everyone would purchase LTC insurance. There is a great deal of evidence that most people understate the probability and the severity of far distanced dependence. This type of myopia or neglect calls for public action. Finally, private LTC insurance is far from actuarially fair; the loading costs are high and lead even farsighted individuals to keep away from private insurance: low income individuals will rely on family solidarity or social assistance and high income individuals on self-insurance.

Cremer and Pestieau (2011) study the role of social LTC insurance in a setting which accounts for the imperfection of income taxation and private insurance markets. Policy instruments include public provision of LTC as well as a subsidy on private insurance. The subsidy scheme may be linear or nonlinear. For the nonlinear part, they look at a society made of three types: poor, middle class and rich. The first type is too poor to provide for dependence; the middle class type purchases private insurance and the high income type is self-insured.³⁷

³⁶ Here, we have in mind disability at very old age.

³⁷ As shown above, this occurs when the marginal utility of LTC is lower than the marginal utility of consumption of healthy elderly.

Two crucial questions are then: (1) at what level should LTC be provided to the poor? and (2) Is it desirable to subsidize private LTC for the middle class? Interestingly, the results are similar under both linear and nonlinear schemes.

First, in both cases, a (marginal) subsidy of private LTC insurance is not desirable. As a matter of fact, private insurance purchases should typically be taxed (at least at the margin). Second, the desirability of public provision of LTC services depends on the way the income tax is restricted. In the linear case, it may be desirable only if no demogrant (uniform lump-sum transfer) is available. In the nonlinear case, public provision is desirable when the income tax is sufficiently restricted. Specifically, this is the case when the income is only subject to a proportional payroll tax while the LTC reimbursement policy can be nonlinear.

Cremer and Roeder (2012) extend this analysis by introducing myopia, and more specifically misperception of LTC risk. They show that social LTC provision is never second-best optimal when private insurance markets are fair. This is perfectly in line with Rochet's results and due to the positive correlation between longevity (and, hence, LTC needs) and productivity. Roughly speaking, the fair private insurance does not redistribute at all while the social insurance redistributes in the "wrong" direction. At the other extreme, when the loading factor in the private sector is sufficiently high, private coverage is completely crowded out by public provision. For intermediate levels of the loading factors, the solution relies on both types of insurance. Regarding private LTC insurance, a myopic agent's tax on private LTC insurance premiums involves a tradeoff between paternalistic and redistributive (incentive) considerations and we may have a tax as well as a subsidy on private LTC insurance.

Means tested long-term care schemes and strategic impoverishment

Public long-term care systems in the OECD are very heterogeneous across and within countries. They vary in generosity, in the levels of government that are involved and in their universality. They are mainly provided by local authorities; they generally only cover a fraction of the needs and range from universal and comprehensive to means-tested systems. In this section, we focus on the means tested systems that seem to prevail in the majority of countries. The best known and the most studied

of them is the Medicaid program in the US, which covers about half of the LTC provision for the American elderly dependents.

Means-testing is rarely a first choice. It is often adopted over universal arrangements because it allows us to devote scarce funds to those who need them the most. The problem is that in reality, needy people do not always have access to means tested programs and well to do individuals can benefit from them. Reasons for this paradoxical outcome can be the fact that the neediest often lack relevant information to take up and fear stigmatization to a larger extent than the members of the middle class. This is particularly true for means-tested public LTC. The reasons are varied. First, there exists a large range of strategies that lead the beneficiary to impoverish himself so as to be eligible. In the US, this is called the Medicaid impoverishment technique.³⁸ Second, most LTC programs seem to favor aid to people who are institutionalized and are unable to meet their financial obligations after a few years. Low-income families are rarely in this situation. Third, there is the practical implementation of means-testing for which the precise definition of “means” is not always clear. Does this concern the income flows or the assets of the beneficiary? Is there a possibility of recouping part of what has been paid by the government at the time of death? Can children be asked to finance their parents’ LTC expenses before the government intervenes? The law is not clear on that. To take the example of France where there are two means tested programs, the PSD (*Prestation Spécifique Dépendance*) and the APA (*Allocation Personnalisée d’Autonomie*), the first can recuperate its participation on the estate of the beneficiary, whereas the second cannot. Finally, and above all, there is a political economy issue. One often has the feeling that there is a political resistance towards implementing means-tested programs when they concern dependent people. When the PSD in France or Medicaid in the US tries to get reimbursed from the estate of a person who has benefited from means-tested services for years, this makes the headlines of newspapers and is perceived as unpopular by the majority of public opinion. In these two countries, as in many others,

³⁸ In the US, there are different strategies for giving away assets to qualify for Medicaid coverage. There exists an army of attorneys who specialize in Medicaid Eligibility and/or Elder Law and can indicate the best way of being eligible and keeping some control of one’s assets. One technique is to transfer one’s assets to an irrevocable trust or to an annuity scheme. A more radical but also riskier approach is to proceed with *inter vivos* gifts to one’s children or grandchildren.

there exist estate recovery programs that are intended to enable states to recoup their expenses upon a beneficiary's death. As it appears, the rate of recovery is extremely low.

Cremer and Pestieau (2012) examine part of these issues. They take a normative viewpoint and cope with the design of a social LTC insurance that would avoid giving away benefits to those who could afford to buy them or receive them from altruistic children. The setting is one of asymmetric information since we know that with perfect information (and absent any political constraints), a first best public LTC scheme could easily be designed and implemented. They analyze two approaches that can be combined. The first relies on a process of self-selection. If private resources cannot supplement the LTC benefit and if its quantity/quality is not very high, those with enough resources or with family support will be deterred from using the means tested scheme. The second relies on explicit control of intergenerational transfers. At some cost, the government can observe *inter vivos* gifts that would be made to impoverish the dependent individual so as to make him eligible for public LTC and enrich his offspring.

They start by looking at an economy where two types of families co-exist. The first type, labeled altruistic, consists of a parent and a child who share the same welfare function. The second type, labeled selfish, consists of a parent and a child with no (financial) links (unlike the altruists, they do not pool their resources). They assume away private insurance for LTC. In a market economy, if the selfish parent is poor, he will be in very bad shape in case of dependence. The altruistic family is assumed to be relatively well off. Whether this comes from the child or from the parent does not matter as they share everything. In case of dependency, the altruistic parent will get a good level of LTC because of his own resources or because of the aid of his child. If the only objective of the government is to make sure that the dependent elderly gets taken care of, it will under perfect information only help the selfish dependent. Assume now that the government does not observe who is altruist and who is not, nor the resources of the parents. The altruistic dependent parent can claim that he is poor and helpless. That way, he gets the public LTC benefit and can either give his assets to his child or, alternatively, he does not have to be helped by his child. How can one avoid this unwanted outcome? First, one can just make sure that the overall level of LTC is

observable and provide an amount of social benefit that is such that the altruistic and wealthy dependent is deterred from mimicking the selfish dependent. The problem with this approach is that it can imply a level of public LTC that is low. An alternative or even a complementary approach is to control both the assets and the gifts of the altruistic dependent through some type of audit. The problem with this approach is that it can be costly.

The political economy of long-term care

Given the large heterogeneity in how LTC expenditures are financed across countries, De Donder and Pestieau (2012) study the determinants of the individual demand and political support for social, private and self-insurance (i.e., saving) in an environment where people differ in income, risk and availability of family help. They start with a setting where only social and self-insurance are available, with social insurance providing a uniform benefit to any dependent person, financed by a proportional payroll tax. The demand for social insurance is shown to decrease with income (because of its redistributiveness across income levels), with family help and to increase with the probability of becoming dependent, when income, risk and family support are independent from each other. Agents with a large income or a very low risk prefer self-insurance (saving) to social insurance. Assuming a positive correlation between income and disability, the relationship between income and the most-preferred social insurance level can go both ways. The correlation between income and family help is not clear. With a positive correlation, richer people unambiguously prefer less social insurance, while the relationship between income and most-preferred social insurance can go both ways with a negative correlation. They show the existence of a majority chosen social insurance level, which decreases with the availability of family help in the economy.

Then, they introduce an actuarially fair private insurance into the picture. The main result is that if agents only differ in income, the introduction of private insurance does not affect its majority chosen level. The intuition is that private insurance induces all above-average-income agents to switch their support in favor of private (rather than social) insurance, but does not affect the preferences of below-average-income agents. With a loading factor, the demand for private insurance decreases

both at the extensive and the intensive margin, up to the point where it becomes nil and where rich agents prefer to exclusively self-insure.³⁹

5. Conclusions

There is a strong feeling that the era of LTC has arrived and represents a crucial challenge for the decades to come. Right now, the provision of LTC is not adequate and the future appears to be gloomy. The source of the problem is twofold, demographic and societal. On the one hand, one witnesses a rapid increase of people aged 80+. The issue of dependency arises precisely in that age bracket. On the other hand, with the drastic change in family values, the increasing number of childless households, the mobility of children and the increasing rate of labor market activity of women, particularly those aged 50-65, the number of dependent elderly who cannot count on the assistance of anyone is likely to increase. Those two parallel evolutions explain why there is a mounting demand on the government and on the market to provide alternatives to the family. But it is not clear that the reasons that explain why the role of the state and the market has been so small up to now will suddenly disappear.

In this paper, we have discussed the nature of these causes and the extent to which we can expect them to fade away. The solution of LTC has to be found in an integrated view of the role of the market, the state and the family. One needs public authorities that are ready to adopt policies that welcome and even foster the intervention of both the market and the family. Solutions exist but they will not bring us to the first best optimum. There are indeed problems that cannot be solved even with the best will. The fact that individuals act opportunistically and that they will then hide both characteristics and actions that can be used by private insurers and the government cannot be avoided. This being said, the tracks of reform are known. First of all, like for the annuity market, much can be done to thicken the LTC insurance market. The government can certainly help but the industry itself has its own responsibility and should in the future exhibit more imagination by offering insurance packages that better fit the

³⁹ See also Nuscheler and Roeder (2010) who study how the heterogeneity in individual income and risk affects the preferences for redistributive income taxation *versus* public financing of LTC.

needs of individuals. Regarding family solidarity, there are measures (part time, tax deduction) that can be taken to facilitate combining work and assistance. It is important to remember that family solidarity is crucial but should rest as much as possible on chosen rather than on forced altruism. Finally, the government can intervene not only indirectly by fostering private insurance and family assistance but directly by providing all sorts of services including social insurance. Above all, a real political will is needed. Even though we are all threatened by dependency, LTC remains an unattractive political issue. We hope that this will soon change.

References

- Adam, S., Bay, C., Bonsang, E., Germai, S. and Perelman, S. (2007), *Retraite, activités non professionnelles et vieillissement cognitif. Une exploration à partir des données de SHARE*, *Economie et Statistiques* No 403-404, 83-96.
- Agarwal, S., Driscoll, J., Gabaix, X. and Laibson, D. (2007), *The age of reason: Financial decisions over the lifecycle*, NBER Working Paper 13191.
- Bernheim, B., Shleifer, A. and Summers, L. (1985), *The strategic bequest motive*, *Journal of Political Economy* 93, 1045-1076.
- Bien, F., Chassagnon, A. and Plisson, M. (2012), *La demande d'assurance dépendance dans un cadre trivarié*, manuscript, PSL, Paris.
- Bisin, A. and Verdier, T. (2001), *The economics of cultural transmission and the dynamics of preferences*, *Journal of Economic Theory* 97, 298-319.
- Brown, J. (2007), *Rational and behavioural perspectives on the role of annuities in retirement planning*, NBER Working Paper 13537.
- Brown, J., Coe, N. and Finkelstein, A. (2007), *Medicaid crowd out of private LTC insurance demand: Evidence from the health and retirement survey*, in J. Poterba (ed.), *Tax Policy and the Economy* 21, MIT Press, Cambridge MA.
- Brown, J. and Finkelstein, A. (2007), *Why is the market for LTC insurance so small?*, *Journal of Public Economics* 91, 1967-1991.
- Brown, J. and Finkelstein, A. (2008), *The interaction of public and private insurance: Medicaid and the LTC insurance market*, *American Economic Review* 98, 1083-1102.
- Brown, J. and Finkelstein, A. (2009), *The private market for long term care in the U.S. A review of the evidence*, *Journal of Risk and Insurance* 76, 5-29.
- Brown, J. and Finkelstein, A. (2011), *Insuring long term care in the U.S.*, *Journal of Economic Perspectives* 25, 119-142.
- Cambois, E., Clavel, A., Romieu, I. and Robine, J-M. (2008), *Trends in disability-free life expectancy at age 65 in France: Consistent and diverging patterns according to the underlying disability measure*, *European Journal of Ageing* 5, 287-298.
- Canta, C. and Pestieau, P. (2012), *Long term care and family norm*, CORE Discussion Paper 2012/22, Université Catholique de Louvain.

- Cawley, J., Grabowski, D. and Hirth, R. (2004), Factor substitution and unobserved factor quality in nursing homes, NBER Working Paper 10465.
- Courbage, C. and Roudaut, N. (2008), Empirical evidence on LTC insurance purchase in France, *Geneva Papers on Risk and Insurance* 33, 645-658.
- Cremer, H., De Donder, P. and Pestieau, P. (2009), Providing sustainable long term care: A looming challenge, TSE Note 3.
- Cremer, H., Gahvari, F. and Pestieau, P. (2012), Uncertain altruism and long term care, manuscript, Toulouse School of Economics.
- Cremer, H., Lozachmeur, J.M. and Pestieau, P. (2012), LTC insurance contract: Lump-sum or cost-sharing?, manuscript, Toulouse School of Economics.
- Cremer, H. and Pestieau, P. (2010), Securing long-term care in the EU: Some key issues, CESifo DICE Report 7(4), 8-11.
- Cremer, H. and Pestieau, P. (2011), Long term care social insurance and redistribution, CORE Discussion Paper 2011/24, Université Catholique de Louvain.
- Cremer, H. and Pestieau, P. (2012), Means-tested LTC and family transfers, manuscript, Université Catholique de Louvain.
- Cremer, H. and Roeder, K. (2012), Long term care policy, myopia and redistribution, manuscript, Université Catholique de Louvain.
- Cutler, D. (1993), Why doesn't the market fully insure long term care?, NBER Working Paper 4301.
- De Donder, P. and Pestieau, P. (2011), Private, social and self-insurance for LTC. A political economy approach, CORE Discussion Paper 2011/53, Université Catholique de Louvain.
- Duée, M., Rebillard, C. and Penneç, S. (2005), Les personnes dépendantes en France: Evolution et prise en charge, XXVème Congrès de l'UIESP.
- EC (2009), The 2009 ageing report, Joint Report prepared by the European Commission (DGECFIN) and the Economic Policy Committee (AWG).
- Eeckhoudt, L., Mahul, O. and Moran, J. (2003), Fixed reimbursement insurance: Basic properties and comparative statics, *Journal of Risk and Insurance* 70, 207-218.
- Finkelstein, A., Luttmer, E. and Notowidigdo, M. (2008), What good is wealth without health? The effect of health on the marginal utility of consumption, NBER Working Paper 14089.
- Finkelstein, A. and McGarry, K. (2006), Multiple dimensions of private information: Evidence from the long-term care insurance market, *American Economic Review* 96, 938-958.
- Finkelstein, A., McGarry, K. and Sufi, A. (2005), Dynamic inefficiencies in insurance markets: Evidence from LTC insurance, *American Economic Review, Papers and Proceedings* 95, 224-228.
- Grabowski, D. and Gruber, J. (2007), Moral hazard in nursing home use, *Journal of Health Economics* 26, 560-577.
- Hiedemann, B., Savinsky, M. and Stern, S. (2012), Will you still want me tomorrow? The dynamics of families long term-care arrangements, manuscript, Université Catholique de Louvain.
- Hiedemann, B. and Stern, S. (1999), Strategic play among family members when making long term care decisions, *Journal of Economic Behaviors and Organization* 40, 29-57.

- Hoerger, T.J., Picone, G. and Sloan, F. (1996), Public subsidies, private provision of care and living arrangements, *Review of Economics and Statistics* 78, 428-440.
- Jousten, A., Lipszyc, B., Marchand, M. and Pestieau, P. (2005), Long-term care insurance and optimal taxation for altruistic children, *FinanzArchiv – Public Finance Analysis* 61, 1-18.
- Kemper, P. and Murtaugh, C.M. (1991), Lifetime use of nursing home care, *New England Journal of Medicine* 324, 595-600.
- Kessler, D. (2007), The long-term care insurance market, *The Geneva Papers on Risk and Insurance – Issues and Practice* 33, 33-40.
- Kopczuk, W. and Slemrod, J. (2005), Denial of death and economic behaviour, *B.E. Journal of Theoretical Economics* 5(1), article 5.
- Lakdawalla, D. and Philipson, T. (2002), The rise in old age longevity and the market for long term care, *American Economic Review* 92, 295-306.
- Meier, V. (1999), Why the young do not buy long term care insurance?, *Journal of Risk and Uncertainty* 8, 83-98.
- Murtaugh, C.M., Kemper, P., Spillman, B.C. and Carlson, B.L. (1997), The amount, distribution and timing of lifetime nursing home use, *Medical Care* 35, 204-218.
- Newhouse, J. (2002), *Pricing the Priceless: A Health Care Conundrum*, MIT Press, Cambridge MA.
- Norton, E. (2000), Long term care, in A. Cuyler and J. Newhouse (eds.), *Handbook of Health Economics* 1b, Elsevier, Amsterdam.
- Nuscheler, R. and Roeder, K. (2010), The political economy of long-term care, manuscript, University of Augsburg.
- OECD (2006), *Projecting OECD health care and long-term care expenditures*, OECD Economics Department Working Paper 477, Paris.
- Pauly, M.V. (1990), The rational non-purchase of long term care insurance, *Journal of Political Economy* 98, 153-168.
- Pauly, M.V. (1996), Almost optimal social insurance of LTC, in R. Eisen and P. Sloan (eds.), *Long Term Care: Economic Issues and Policy Solutions*, Kluwer, London.
- Pèrès, K., Jagger, C., Lièvre, A., and Barberger-Gateau, P. (2005), Disability-free life expectancy of older French people: Gender and education differentials from the PAQUID cohort, *European Journal of Ageing* 2, 225-233.
- Pestieau, P. and Ponthiere, G. (2011), The long term care insurance puzzle, in J. Costa-Font and C. Courbage (eds.), *Financing Long term Care in Europe: Institutions, Markets and Models*, Palgrave Macmillan, Basingstoke.
- Pestieau, P. and Sato, M. (2006), Long term care: The state and the family, *Annales d'Economie et de Statistique* 83/84, 123-150.
- Pestieau, P. and Sato, M. (2008), Long term care: The state, the market and the family, *Economica* 75, 435-454.
- Pezzin, L., Pollak, R. and Schone, B. (2007), Efficiency in family bargaining: Living arrangements and caregiving decisions of adult children and disabled elderly parents, *CESifo Economic Studies* 53, 69-96.
- Pezzin, L., Pollak, R. and Schone, B. (2009), Long term care of the disabled elderly: Do children increase caregiving by spouses?, *Review of Economics of the Household* 7, 323-339.
- Ponthiere, G. (2011), Long term care, altruism and socialization, PSE Discussion paper hal-00622385, Paris.

- Rivlin, A., Weiner, J., Hanley, R. and Spence, D. (1988), *Caring for the Disabled Elderly: Who will Pay?*, The Brookings Institution, Washington DC.
- Rochet, J.C. (1991), Incentives, redistribution and social insurance, *The Geneva Risk and Insurance Review*, Palgrave MacMillan 16, 143-165.
- Sloan, F., Hoerger, T. and Picone, G. (1996), Effects of strategic behaviour and public subsidies on families' savings and long-term care decisions, in R. Eisen and F. Sloan (eds.), *Long-Term Care: Economic Issues and Policy Solutions*, Kluwer Academic Publishers, Boston.
- Sloan, F. and Norton, E. (1997), Adverse selection, bequests, crowding out, and private demand for insurance: Evidence from the LTC insurance market, *Journal of Risk and Uncertainty* 15, 201-219.
- Stern, S. and Engers, M. (2002), LTC and family bargaining, *International Economic Review* 43, 73-114.
- Taleyson, L. (2003): *L'assurance dépendance privée: Comparaisons internationales*, Newsletters techniques SCOR.

Comment on Cremer, Pestieau and Ponthiere: The economics of long-term care: A survey

Þórólfur Matthíasson*

A person in need of long-term care (LTC) is a person that is unable to carry out daily activities without assistance supplied by someone else. Care given to infants is not classified as long-term care, though. Hence, LTC is required by the elderly and the disabled. The authors restrict their survey to papers dealing with LTC required by the elderly.

1. The market for long term care is not very orderly

Total demand is not so much a function of price as it is a function of things like the share of elderly in the population, the incidence of disability among the elderly and the length of time they are in need of assistance. Increasing or decreasing the price of LTC will not affect these issues in any significant way. Given the age distribution of the population in most developed economies, it is easy to postulate that demand will increase: the share of elderly is increasing, longevity is also increasing so that the demand for LTC will increase even if the length of time each person needs assistance is fixed. This is surely and squarely pointed out by Cremer, Pestieau and Ponthiere.

The supply of long-term care comes from many sources: Spouses, parents, children, grandchildren, stepchildren, the state, local communi-

* Faculty of Economics, University of Iceland, totimatt@hi.is.

ties, charities, non-profit organizations, private for profit organizations to name but a few.

Supply is driven by love, by thankfulness for care given by the receiver to the caregiver at an earlier point in time, as an act of duty, in hope of a reward in the form of a big inheritance. Sometimes supply is driven by payment and profit. Hence, a higher price for LTC will to some extent increase the supply of LTC. But it is also highly probable that a higher price of LTC will shift supply from being a not-for-profit to a for-profit activity. Thus, it is possible, but not probable, that a higher price of LTC results in a lower total supply of LTC!

It is hard to come up with a comprehensive model that takes all aspects of supply and demand for long-term care into consideration. In particular when one takes into consideration that it is likely that the structure of the market is going to change in fundamental ways in the next decades in many of the most developed countries of the world.

In the introduction, the authors point out that the market for long-term care is almost non-existing except in the US and France. Most of the care is given under some kind of a family umbrella.

The authors introduce the long-term care insurance-market puzzle. While the cost of long-term care runs at forty thousand euros in France per dependent per year and forty to seventy thousand dollars in the US, very few people insure against these costs. The expected dependency time is some two to three years, so here we have a risk that runs in hundreds of thousands of euros or dollars that every man and every woman are facing and very few insure against it.

The authors give a good survey of attempts to explain the puzzle and I will not attempt to retell their story, but part of the story is of course that there are implicit insurance contracts within families, local communities and between the individual and the state.

But part of the story is that most people in northern Europe at least assume that they are insured anyway: We assume that as part of our tax-money goes to pay for long term-care of the elderly now, we have a claim in the future on those that are taxpayers then to pay for some of the cost of our expected long-term care costs. We assume that insuring will only result in a situation where we end up paying for something that we would have gotten free of charge anyway.

The authors give a good overview of recent works aimed at explaining how a rational individual might go about increasing the likelihood of being taken care of late in life by relatives. The keywords are: Family games, strategic bequest motives. Those are games that people play and it is natural that economists try to model the games and deduce how they affect the world we live in. But it is hard not to smile from time to time. The assumption in the models is that people start to motivate their children relatively late in life. We should be able to do better than that and try to model decisions to marry, to have children, decisions regarding the number of children etc. Well, here the reader who is or has been married can use some introspection: Did you think of long-term care at old age when deciding for or against marriage?

Anyhow, this reviewer is of the opinion that the authors have managed to show us that there is some work to be done in this field.

What I did miss was more emphasis on the hidden logic of an aging society. Questions like: How is it possible to deliver long-term care if every person of old age is insured and the share of old people in the population approaches 1? When would you have to start building up funds abroad in order to tally such a task? How large should such a fund be?

Finally, the conflict between individuals in need of long term care and their relatives and potential heirs is not new. Egill SkallaGrímsson the Viking poet spent his last year at his farm Mosfell close to Reykjavik. He was a wealthy man and wanted to use some of his wealth to have fun during his final years. In particular he wanted to take his money to Althingi while Althingi was in session in Thingvellir and throw out the coins among the nobilities. The amusement was to be to see how greed would diminish those attending the parliament. Egill's relatives did not see this as a wise use of funds and denied to help him in this respect. Hence, Egill used the opportunity when left at home alone with two slaves to have them help him bury the treasure. To ensure that no one would ever find the treasure, he killed both slaves in return for their service!

The moral of the story of Egill and his treasure is that one can be forced to spend money saved for amusement late in life on very different projects.

The role of primary health care in controlling the cost of specialist health care^{*}

Stephen Beales^{**} and Peter C. Smith^{***}

Summary

There is some concern that expenditure levels in developed health systems are reaching unsustainable levels. One cause appears to be the ‘avoidable’ use of expensive specialist services. The belief is that – with timely, high quality primary care – such utilization could be markedly reduced, with associated cost savings and improved quality of life for patients. This paper reviews the evidence for three broad forms of primary care intervention: reducing or delaying the onset of disease; reducing the use of specialist care once a clinical condition has been identified; and reducing the intensity of specialist care once a need for such care has arisen. We examine the role of incentives in promoting the cost containment role of primary care. The paper concludes with a discussion of the associated policy implications.

Keywords: primary healthcare, specialist healthcare, cost saving, efficiency, chronic healthcare, disease prevention, incentives.

JEL classification numbers: I11; I18.

^{*} The authors are grateful for comments from the editors Tor Iversen and Sverre Kittelsen, the discussant Helgi Tómasson, and other participants at the Reykjavik seminar at which an earlier draft of this paper was presented.

^{**} Centre for Health Policy, Imperial College London, s.beales@imperial.ac.uk.

^{***} Centre for Health Policy, Imperial College London, peter.smith@imperial.ac.uk.

Primary care, and specifically general medical practitioners, play a major role in many health systems. In such systems, the general practitioner (GP) or family physician has the prime responsibility for providing comprehensive health care and arranging for other health personnel to provide services when necessary. The general practitioner functions as a generalist who accepts everyone seeking care. The Declaration of Alma-Ata, made in 1978, became a core concept in the World Health Organization's goal of 'Health for All'. The Declaration places great importance on primary health care and it states: "It [primary health care] forms an integral part both of the country's health system, of which it is the central function and main focus, and of the overall social and economic development of the community."

Advocates for the importance of primary care within the health system claim that – effectively organized – it can both improve the health of the population and reduce the aggregate costs of health care (Starfield, 1992; Macinko et al., 2003). According to this argument, the organization of primary care might be a crucial influence on both the effectiveness and the cost-effectiveness of the health system. The Declaration of Alma-Ata defines primary care as "essential health care based on practical, scientifically sound and socially acceptable methods and technology made universally accessible to individuals and families in the community through their full participation and at a cost that the community and the country can afford to maintain at every stage of their development in the spirit of self-reliance and self-determination".

Looking across 12 (western industrialised) countries during the mid-to-late 1980's, Starfield (1992) showed that countries with a stronger orientation towards primary care were more likely to have better health levels and lower costs. Her score for the strength of primary care orientation was based on five health system characteristics and six practice characteristics. The five system characteristics were the extent to which health professionals and facilities were regulated so that they were geographically distributed approximately according to need; the type of physician designated as the primary care physician; the professional earnings of primary care physicians relative to other specialists; the number of primary care physicians relative to other specialists; and the extent of insurance coverage for health services. The six practice characteristics were the extent to which people sought care from their primary care physician

before going elsewhere; the strength of relationships between people and their primary care physician; the extent to which primary care practice dealt with common needs regardless of their type; the degree of coordination between primary care and other health services; family orientation of primary care; and community orientation of primary care. This study suggests that, as well as directing resources towards those who need them most and not excluding some from coverage, ensuring that people visit primary care first reduces the costs and improves the overall quality within the system. The characteristics that the study looked at were broad and the extent to which features within these characteristics made a greater or smaller difference to the overall cost and quality is unclear. There could also be a self-selection bias, as countries that were already more cost conscientious might choose particular characteristics, such as gate-keeping, rather than that characteristic causing the cost reduction.

Amongst the many roles of primary care, one of the most prominent is maintaining the health of the populations, and in particular minimizing the avoidable use of expensive secondary and tertiary care. Success in this is a signal of good population health, but is also likely to contribute to health system expenditure control. In particular, there is a group of conditions known as ambulatory care sensitive conditions (ACSCs) for which emergency admissions into hospital are thought to be avoidable through better primary care. It is being recognized to an increasing extent that reducing hospital admissions for ACSCs may be a key requirement for maintaining the financial sustainability of publicly funded health services. Table 1 summarizes the rates of potentially avoidable emergency admissions to hospitals in OECD countries for three common chronic conditions: asthma, chronic obstructive pulmonary disease (COPD) and diabetes, highlighting very large apparent variations between countries.

Table 1. Hospital admission rates per 100 000 population aged 15 or above, age/sex standardized, 2009 or nearest year

	Asthma	COPD	Diabetes
Australia	66.6	311.7	7.5
Austria	52.8	310.1	187.9
Belgium	48.4	227.6	-
Canada	15.7	183.3	15.2
Czech Republic	37.0	149.5	31.4
Denmark	36.5	276.8	65.4
Finland	75.9	146.5	78.3
France	43.4	79.1	-
Germany	20.8	200.6	50.3
Hungary	35.0	247.8	129.2
Iceland	33.3	229.4	20.4
Ireland	43.5	363.9	32.4
Israel	68.4	233.5	7.0
Italy	19.2	126.2	33.1
Korea	101.5	221.9	127.5
Latvia	120.7	163.0	17.5
Malta	78.9	134.8	40.8
Mexico	19.0	111.4	108.9
Netherlands	27.5	154.4	-
New Zealand	80.7	319.5	7.6
Norway	47.6	243.0	46.7
Poland	68.9	216.8	65.9
Portugal	15.1	71.3	16.3
Singapore	86.4	183.0	18.5
Slovak Republic	166.8	206.2	-
Slovenia	38.1	113.7	42.0
Spain	43.9	139.3	3.3
Sweden	19.3	137.5	66.0
Switzerland	30.9	91.5	18.8
United Kingdom	73.7	213.4	23.9
United States	120.6	229.8	21.2
OECD	52.7	198.8	41.4

Source: OECD Health Data 2011.

Table 2. Extent and nature of gatekeeping in OECD countries

Country	Does every citizen have to register with a primary care physician?	Does access to secondary care require a GP referral?
Australia	No obligation and no incentive	Financially encouraged
Austria	No obligation and no incentive	No obligation and no incentive
Belgium	Financially encouraged	Financially encouraged
Canada	No obligation and no incentive	Compulsory
Czech Republic	No obligation and no incentive	No obligation and no incentive
Denmark	Compulsory	Compulsory
Finland	No obligation and no incentive	Compulsory
France	Financially encouraged	Financially encouraged
Germany	Financially encouraged	Financially encouraged
Greece	No obligation and no incentive	No obligation and no incentive
Hungary	Financially encouraged	Compulsory
Iceland	No obligation and no incentive	No obligation and no incentive
Ireland	No obligation and no incentive	Financially encouraged
Italy	Compulsory	Compulsory
Japan	No obligation and no incentive	No obligation and no incentive
Korea	No obligation and no incentive	No obligation and no incentive
Luxembourg	No obligation and no incentive	No obligation and no incentive
Mexico	No obligation and no incentive	Compulsory
Netherlands	Compulsory	Compulsory
New Zealand	Financially encouraged	Compulsory
Norway	Compulsory	Compulsory
Poland	No obligation and no incentive	Compulsory
Portugal	Compulsory	Compulsory
Slovak Republic	Compulsory	Compulsory
Spain	Compulsory	Compulsory
Sweden	No obligation and no incentive	No obligation and no incentive
Switzerland	Financially encouraged	Financially encouraged
Turkey	No obligation and no incentive	No obligation and no incentive
United Kingdom	Financially encouraged	Compulsory

Source: OECD Survey on health system characteristics 2008-2009 <http://dx.doi.org/10.1787/810665718628>.

In examining the role of GPs, a particularly important debate concerns the function of ‘gatekeeping’, under which a patient may secure access to non-emergency care, prescription drugs and other health services only with the authorization of a GP. This gatekeeping role has been a central feature of the UK National Health Service since its inception in 1948, but also plays an important part in many other publicly financed health systems, in particular in some Scandinavian systems, and even certain US

managed care organizations. Recent reforms in traditional social health insurance systems, such as Germany and France, have included tentative steps towards a stronger gatekeeping function. Several objectives of gatekeeping have been proposed, but prominent amongst them has been the desire to constrain the natural tendency of patients to use health services in excess of socially optimal levels when they do not bear the full marginal cost of treatment. Table 2 summarizes the nature and extent of gatekeeping in OECD countries.

No discussion of the role of primary care is possible without considering the incentives under which it functions. These incentives might be directly financial, in the form of payment mechanisms, or non-financial, for example in the form of performance report cards and the implied implications for reputation. Financial incentives arise from the payment mechanisms under which primary care is reimbursed. Depending on how these are structured, they have the potential either to increase or decrease system efficiency. For example, if they directly reward evidence-based preventative medicine and other actions designed to reduce future demands for health services, they may serve to promote the cost-effectiveness of the health system. If, on the other hand, they offer no reward for timely intervention and mitigation of future ill-health, the financial incentives may be dysfunctional and serve to exacerbate cost pressures, for example by encouraging referral for unnecessary specialist diagnostic tests, or indeed by discouraging necessary referrals and thereby increasing future healthcare costs. In the same vein, non-financial instruments, most notably performance reporting requirements, might reinforce or work against cost containment objectives, depending upon their design.

This paper examines the published evidence on the role of primary health care in controlling the cost of specialist health care. The next section examines three fundamental functions of primary care that may contribute to this objective: reducing or delaying the onset of disease; reducing the use of specialist care once a clinical condition has been identified; and reducing the intensity of use of specialist care once a need for such care has arisen. The following section examines the role of incentives in promoting the cost containment role of primary care. A concluding section discusses the broader alignment of health system design with that objective, and the policy implications.

1. How might primary care reduce the costs of specialist care?

Primary care might be expected to reduce the costs in secondary care in three broad ways:

- Improving the health of the population, thereby reducing the incidence and severity of disease, and the associated need for specialist care. In this domain, the prime role of primary care might take the form of conventional disease prevention activities, such as vaccination programmes, or influencing behavioural change aimed at reducing the risk of future disease.
- Reducing the use made of secondary care once a clinical condition has been identified, usually in the form of chronic disease (for instance, reducing emergency admissions for diabetes complications). Many of the activities in this area can be considered under the generic heading of ‘disease management’.
- Reducing the intensity of any specialist utilization once a need for specialist care has been identified. This role can be considered under the general heading of care integration, and may take the form of task substitution (performing some specialist tasks in a lower cost primary care setting) or patient monitoring (allowing earlier discharge of a patient from specialist care).

We consider each of these in turn.

1.1 Improving the health of the population

There is a long-standing belief that, when effectively used, primary care can improve the health of the population, and prevent, mitigate or defer the development of health problems that would otherwise give rise to costly use of health services, including specialist care. Relevant actions might take the form of preventative medicine (immunization, vaccination etc.), behavioural advice (on smoking, alcohol, diet etc.), and timely intervention when there are early indications of disease (from screening or other diagnostic instruments).

The most rudimentary role of primary care in all health systems is to ensure that children and other vulnerable people are properly vaccinated and immunized against infectious disease. Most developed health systems have in place arrangements to ensure that high levels of such interventions are secured, particularly amongst children.

There is less uniformity in approaches to interventions aimed at reducing actions that are considered to be 'risk factors' for future disease, such as smoking tobacco and alcohol consumption. These interventions are well suited to a primary care setting and, if effectively undertaken, are likely to have an impact on the number of admissions into secondary care.

For example, the National Institute for Health and Clinical Excellence (NICE) in England recommends that a brief intervention, in the primary care setting, is carried out for each patient, referring those that smoke to a smoking cessation program (National Institute for Health and Clinical Excellence, 2006). The intervention is deemed to be cost-effective. An area with an adult population of 131 000 could expect to incur additional costs of GBP 18 000, with an 11-year saving for acute myocardial infarction and stroke of GBP 210 000. The estimate is based on the work of Naidoo et al. (2000) who modelled the expected impact of the intervention by altering the survival probability of those it would affect over time.

Godtfredsen, Vestbo et al. (2002) looked at the effect of a Danish smoking cessation program on chronic obstructive pulmonary disease (COPD) admissions. They found that quitting smoking was associated with a significant reduction in the risk of hospital admission. They noted that those who reduced smoking did not show any significantly lower risk of hospitalisation than continuing heavy smokers.

Although smoking cessation interventions carried out in a primary care or a community setting are known to be cost effective relative to other medical interventions (Fiore et al., 2000) a study in 2001 found that only 21 percent of patient visits to their primary care physician included a discussion about tobacco use (Ellerbeck et al., 2001). Of those identified as smokers, 38 per cent were provided with smoking cessation assistance.

NICE recommends that to tackle alcohol abuse, an opportunistic screening of adults should take place in primary care. Unlike organised screening, in which sections of a population are identified, contacted and offered a screen, opportunistic screening takes place on an ad-hoc basis – in this case whenever someone visits their GP and the GP deems it neces-

sary. Those that are found to be harmful or hazardous drinkers would then be offered a further session of brief advice on how to reduce their alcohol consumption to a safer level. An economic analysis confirmed that several examples of providing screening and brief advice are estimated as providing cost savings (providing additional healthcare benefits and an overall reduced health service cost) (Purshouse et al., 2009). The analysis indicated that NHS and personal social services savings over GBP 124.3 million may be realisable over a 30-year time horizon.

Interventions aimed at securing behavioural change are in general at an early stage of development. They may become crucial to the future financial sustainability of publicly funded health systems, and it is likely that primary care will play a central role in delivering such interventions. However, although successful behavioural interventions will unambiguously deliver improvements in the length and quality of life, there remains a question over the sign and magnitude of their long-term financial impact on the health system. They may lead to a longer lived population that places new demands on the health system, in the form of long-term care for example. There is a clear need for an increased use of population microsimulation methods to assess the long-term impact on the health system of public health interventions (Zucchelli et al., 2010).

1.2 Reducing the use made of secondary care

Once a health need has arisen, there might be an important role for primary care in its treatment and management, thereby affecting costs. Examples might include treatment by primary care of minor trauma or other interventions that would otherwise require a more costly secondary treatment, or the effective management by primary care of established chronic disease (such as diabetes or hypertension), thereby reducing the emergency use of specialist care.

A lack of primary care provision might be one cause of increased secondary care usage. Parchman and Culler (1999) found, when looking at Medicare beneficiaries in fair and poor health in the US, that they were 1.7 times more likely to experience a preventable hospitalisation if they resided in a primary care health professional shortage area, after controlling for educational level, income and supplemental insurance. Similarly, Macinko et al. (2010) showed that, in Brazil, high enrolment in its Family

Health Program was associated with 13 percent lower hospitalisation rates for ambulatory care-sensitive chronic diseases, than in municipalities with low enrolment, when other factors were held constant.

Many acute minor problems in clinical practice are dealt with in hospital accident and emergency (A&E) departments without referral from general practice, even though they might readily be adequately managed by the primary health care team. Problems for which patients use A&E which are thought to be more suitable for assessment and management by a GP include minor injuries and lacerations, acute infections, most eye problems and bites (Myers, 1982).

In the UK, for example, as many as a quarter of those attending A&E departments have minor injuries or illnesses not requiring specialist attention (Lowy et al., 1994). Therefore, there is a large scope for potential savings from providing care for these patients. From a cost-effectiveness perspective, the key question is what, if any, sacrifice in clinical quality arises.

Roberts and Mays (1998) looked at three controlled trials where general practitioners were being based within A&E departments in inner city locations (King's College Hospital and St Mary's Hospital in London and St James' Hospital in Dublin). They found a lower general use of diagnostic investigations by the general practitioners and fewer referrals to secondary services. The King's College study found a more marked difference in the usage of X-rays and, unlike the Dublin study, also found that hospital doctors were more likely to prescribe than general practitioners. There was no evidence of any significant difference in patient satisfaction or health outcomes between general practitioner and hospital doctor management. Both the King's College hospital study and the Dublin study found that employing general practitioners resulted in cost savings.

The role of primary care is likely to be most important in the area of chronic disease. One of the key challenges for health systems in the developed world (and many in the developing world) is the rising prevalence of non-communicable diseases, associated with an ageing population and increases in risk factors such as obesity. Disease management programmes are increasingly being implemented in healthcare systems worldwide in order to enhance the quality and reduce the cost of caring for those with chronic illnesses. Mattke et al. (2007) define disease man-

agement as a patient centred approach of coordinated multiple healthcare interventions that structure chronic care to a specific patient group. One influential model of disease management is the Chronic Care Model (CCM) developed in the US by Wagner, Austin et al. (2001) (see Box 1 for a breakdown of its components).

The components of the CCM model listed in Box 1 are intended to help with the re-design of primary care to improve its care for patients with chronic conditions. A systematic review evaluated the impact of disease management programmes that contained two or more of those CCM components for diabetes, depression, heart failure and COPD (de Bruin et al., 2011). It found that, of the studies that reported changes in healthcare costs, 13 out of 21 showed a decrease in overall costs. The overall costs varied between –USD 16 996 (a system cost saving) and USD 3 305 (an increase in the system cost) per patient per year (in 2007 prices). The results suggest that the results are most positive for disease management programmes for patients with heart failure and least positive for patients with depression. The study found no correlation between particular components of the programme, or the number of components, and the overall cost saving associated with the programme. It concluded that the evidence for proving that disease management programmes saved money was still inconclusive and that further economic evaluations were required.

Other studies have looked at specific interventions that have been introduced by primary care to help improve care for those with chronic diseases and, ultimately, reduce the cost to the system of the patient. Here we look at some of the key interventions that have been found to be cost saving.

Box 1. The chronic care model

In an effort to create a framework that primary care could use to improve the delivery of interventions that would improve the care of those with chronic conditions in America, the Chronic Care Model (CCM) was developed by Wagner, Austin et al. (2001).

According to Bodenheimer et al. (2002), there are 6 key components to the CCM framework:

1. Health care organization

The reimbursement environment of a provider organization has a major impact on chronic care improvements, which are more likely to survive in the long-term if they increase revenues or reduce expenses. If purchasers and insurers fail to reward chronic care quality, improvements are difficult to sustain.

2. Community resources and policies

The longer time horizon and the fluctuating course of many chronic illnesses require a regular interaction between caregivers and patients. To improve chronic care, provider organizations need linkages with community-based resources, e.g. exercise programmes, senior centres, and self-help groups, as they are able to provide more frequent support.

3. Self-management support

For chronic conditions, patients themselves become the principal caregivers. People live with chronic illness for many years; management of these illnesses can be taught to most patients, and substantial segments of that management—diet, exercise, self-measurement (e.g. using glucometers or bathroom scales) and medication use—are under the direct control of the patient.

4. Delivery system design

The structure of medical practice must be altered, creating practice teams with a clear division of labour and separating acute care from the planned management of chronic conditions. Physicians treat patients with acute problems, intervene in stubbornly difficult chronic cases, and train team

members. Non-physician personnel are trained to support patient self-management, arrange for routine periodic tasks (e.g. laboratory tests for diabetic patients, eye examinations, and foot examinations) and ensure appropriate follow-ups. Planned visits are an important feature of practice redesign.

5. Decision support

Evidence-based clinical practice guidelines provide standards for optimal chronic care and should be integrated into daily practice through reminders. Ideally, specialist expertise is a telephone call away and does not always require full specialty referral. Guidelines are reinforced by physician “champions” leading educational sessions for practice teams.

6. Clinical information systems

Computerized information has 3 important roles: First as reminder systems that help primary care teams comply with practice guidelines; second as feedback to physicians, showing how each is performing on chronic illness measures such as HbA1c and lipid levels; and third as registries for planning individual patient care and conducting population-based care.

Patient Education

In a study of diabetic patients in rural Austria participating in a structured education and treatment programme, Pieber et al. (1995) found that programme patients had lower health care costs after six months as compared to a control group.

At a staff-model health maintenance organization (HMO), children with asthma were offered a single educational session. The group showed a 40 per cent reduction in emergency department visits (Greineder et al., 1999).

Whereas traditional patient education offers information and technical skills, self-management education teaches problem-solving skills. A central concept in self-management is self-efficacy – confidence to carry out behaviour necessary to reach a desired goal. Self-efficacy is enhanced when patients succeed in solving patient-identified problems. Evidence from controlled clinical trials suggests that programmes teaching self-management skills are more effective – for arthritis and adult asthma –

than information on patient education in improving clinical outcomes and, in some circumstances, self-management education improves outcomes and can reduce costs (Bodenheimer et al., 2002a).

Rich et al. (1995) demonstrated in a randomised controlled trial that a nurse-directed programme of patient education with post-hospital telephone and home visit follow-up (self-management support combined with delivery system redesign in the CCM terminology) was associated with a 56 per cent reduction in hospital readmissions for congestive heart failure (CHF) and a significant improvement in quality-of-life scores as compared with controls. Within a 90-day period, the overall cost of care was USD 460 less per patient in the treatment group. Stewart et al. (1999) showed that using a similar, but less intensive intervention led to a 52 per cent reduction in hospital costs for the intervention group during an 18-month period, resulting in overall savings. Cline et al. (1998) found similar cost reductions with a programme of self-management support and a nurse directed CHF clinic.

Bolton et al. (1991) demonstrated that adult asthma patients who were given self-management support made fewer emergency department visits throughout a 12-month period as compared to a usual care group, with the USD 85-per-person cost for the educational sessions offset by the USD 628-per-person reduction in emergency department charges.

Compared with usual care, a home-based health education programme for low-income children previously hospitalised with asthma saved USD 11 for each dollar spent to deliver the health education (Clark et al., 1986). For children without a recent hospitalisation, the costs for the two groups were the same.

However, in three other studies (Bailey et al. 1990; Wilson et al., 1993; Vojta et al., 1999), self-management support interventions did reduce health care use, but the control groups showed similar reductions. In a study with a longer follow-up period, Kauppinen et al. (2001) found that there was no significant reduction in health care costs after five years for adult asthma patients given intensive patient education throughout a one-year period. And Lahdensuo et al. (1998) found that for patients with mild to moderately severe asthma, the health care costs were actually higher for the group receiving a self-management programme than for a usual care group.

This last result illustrates the challenges associated with patient education programmes. They are likely to increase patient awareness and an improved health status. However, this may also lead to increased healthcare utilization, at least in the short and medium term. The question for evaluation is the extent to which longer term specialist costs are averted by the intervention, an area that is by its nature difficult to assess empirically.

Enhanced Access

One of the reasons for unnecessary use of specialist emergency care is the poor access to primary care, especially out of normal office hours. In Australia, Hamiton (2007) found that patients report that GP accessibility is by far the strongest factor in the decision to attend the emergency department rather than a GP practice. Increased access to GP services may therefore prevent some of these patients from attending the emergency department.

However, the evidence is again equivocal. In a study of CHF patients discharged from nine Veterans Affairs hospitals, Oddone et al. (1999) compared usual care with an intervention including patient education, nurse telephone follow-up and enhanced access to primary care. The number of hospital readmissions did not differ between the two groups and the number of outpatient visits was actually higher in the intervention group.

Cluster visits

An important constraint to the delivery of complex healthcare, especially amongst elderly patients, is the need to coordinate consultations with a range of healthcare professionals. Care may be suboptimal, and resources wasted, if patients are unable or unwilling to make all necessary visits. Primary care is the obvious setting in which such coordination can take place. Kaiser Permanente studied the impact of co-locating a multidisciplinary team so that each relevant member can meet the patient during a single visit (Sadur et al., 1999). The six-month programme for diabetes care resulted in reduced hospital and outpatient use as compared to usual care. The multidisciplinary outpatient diabetes care management was delivered by a diabetes nurse educator, a psychologist, a nutritionist and a pharmacist in cluster visits.

However, a study looking at a similar setup for CHF patients found that it returned no overall cost reductions as compared to usual care (Riegel et al., 2000). Interestingly, although the intervention led to increased costs for Class I patients (those with the mildest symptoms), it did create savings for Class II patients (those with moderate symptoms). This suggests that a better patient selection may make this a more cost-effective intervention.

Case management

Primary care is also an obvious setting in which the ‘case management’ of patients with complex medical needs can be organized, with the intention of improving the health status and reducing the need for emergency care. At a staff-model health maintenance organization (HMO), children with asthma were enrolled in a nurse case-management programme. The group had 73 per cent fewer emergency department visits and 84 per cent fewer hospitalisations, with the savings greatly exceeding the programme costs (Greineder et al., 1999).

Stewart et al. (2002) calculated that it would be cost-effective to introduce a specialist nurse-mediated, post-discharge management service for heart failure for the whole population (of the UK). The authors suggest that such a service would not only improve quality of life and reduce admissions in patients with congestive heart failure, but also reduce costs.

Domurat et al. (1999) looked at an intervention aimed at high-risk diabetes patients from Kaiser Permanente in which the patients were intensively managed by a team that offered planned diabetes visits, telephone contacts and group educational sessions. The study found that patients in the intensive programme remained half as long in hospital as those in the control group. One of the findings of the study was that patients discharged from the intensive-management programme may revert to their pre-programme status, which suggests that ongoing case management is needed, and that the continuing costs need to be balanced against the savings elsewhere.

Johansson et al. (2001) evaluated the effect of an individual support intervention on the utilization of specialist care among cancer patients. They found that the intervention reduced the number of admissions and the (length of stay) LOS after adjustment for weight loss and psychological distress, but only for older patients (those aged 70 or older). The in-

tervention included intensified primary healthcare comprising of extended information from specialist clinics, education and supervision in cancer care for general practitioners, and home-care nurses.

Improved glycaemic control

Improved glycaemic control is seen as a core component of managing diabetes patients. In contrast with programmes for CHF and asthma, which might be expected to produce cost savings almost immediately, through reduced hospital and emergency department use, programmes that improve diabetic glycaemic control might be expected to show savings only in the long term, with reduced vascular complications. However, some studies have shown that improved diabetes care can lead to cost reductions even in the short run. Wagner, Sandhu et al. (2001) compared two groups of diabetic patients throughout one year and found that the organization was saving between USD 685 and USD 950 per patient annually for the group with improved HbA1c levels. The savings resulted from fewer hospital admissions, emergency department visits and physician consultations. The savings were only statistically significant for patients in the improved group whose baseline HbA1c level was 10 percent or above. Measuring HbA1c is a way of measuring blood glucose levels – an important indicator for diabetics. For non-diabetics, a usual reading would be in the range 4.0-5.9 percent, for people with diabetes an HbA1c level of 6.5 percent is considered to be good control. However, this study did not take into account the cost of the intervention and, therefore, the overall cost-effectiveness could not be assessed. In a similar vein, Testa et al. (1998) showed that an improved glycaemic control of type 2 diabetes was associated with short-term reductions in hospital stay.

Pharmacist provided patient education and monitoring

Although there is little empirical evidence, it may be the case that pharmacists can fulfil some of the primary care roles described above. Munroe et al. (1997) found that the total health care costs dropped for diabetic patients enrolled in a programme in which specially trained pharmacists provided patient education, monitoring, and feedback to physicians, when compared with a control group.

1.3 Reducing the intensity of specialist utilization

Once access to specialist care has been secured, the intensity of use, and therefore the costs of that care, might be highly dependent on the organization and capacity of primary care. The cost-effectiveness for an entire episode of care will often depend on the integration of specialist, primary and social care. In particular, with a well-functioning system of primary care, it might be feasible to discharge a patient from hospital sooner than would otherwise be the case, without any increased risk to the patient's health. To this end, some aspects of rehabilitation might be undertaken in a community rather than a hospital setting, and primary care might implement monitoring systems that increase the potential for telemedicine and initiatives such as 'hospital at home'.

A systematic review found that the effectiveness of simple outreach programmes (where specialists see patients in a primary care setting) was uncertain (Bazian, 2005). However, 'enhanced' outreach (entailing an intense involvement of specialists in primary care) was associated with improvements in the appropriateness of care, symptoms, patient satisfaction and concordance. The study could not identify which elements of the intervention caused the improvements in care. In particular, the mental health studies identified in the review examined a wide-ranging and intensive attempt to improve care. The shift of the psychiatrist from the hospital to the community was only one part of the intervention, and was not considered to be the most important by the investigators. The authors suggest that it may be the case that a thorough programme of patient and primary care clinician education, without specialist clinical contact, would have had similar benefits and that this approach would still enable primary care clinicians to provide appropriate care for a wider range of patients, some of whom might otherwise have been referred to secondary care, without compromising the availability of clinicians in secondary settings. They concluded that this approach might cost less, be more efficient, and lead to better care than full outreach. The study did not, however, look at system cost and the authors do not speculate on whether they believe that the interventions saved money overall.

A similar scheme, one of a number called Hospital at Home, sought to reduce the number of people requiring inpatient care by treating them at home, and was found to save expenditure (Jones et al., 1999). Hospital at

home can provide an alternative to inpatient care in two ways – early discharge of patients from hospital or avoidance of admission.

Patel et al. (2004) evaluated the costs of stroke care in three different settings – stroke unit, stroke team and domiciliary stroke care – and found that over a twelve-month period domiciliary care was the cheapest setting. The incremental cost per quality-adjusted life year gained was GDP 64 097 between domiciliary care and stroke unit care, leading the authors to conclude that cost perspectives are important when stroke services are evaluated and that the improved health outcomes in the stroke unit come at a higher cost (currently in excess of the usual threshold for accepting new technologies in England).

One reason for unnecessary specialist costs may be that a patient's home circumstances make early discharge from hospital infeasible (for example, if there is no caregiver at home), leading to so-called 'blocked' hospital beds. This outcome is particularly likely to arise in systems where hospital and social care are delivered by separate agencies. In England and Wales, the health services have the power to fine local governments a daily tariff for delays in discharge caused by local social care failures. An alternative model was that local hospital and social care services were encouraged to work collaboratively by applying for special grants to improve community services. A study in 2007 found that almost two thirds of the hospitals opted not to charge the local governments, but instead opted to work collaboratively (McCoy et al., 2007). The study found no improvement in delayed discharge bed days, beyond what the authors deemed to be a long-term trend of reduction.

Holmas et al. (2010) studied the effects of fining owners of long-term care institutions in Norway who prolong hospital LOS, driving hospital costs upwards and causing bed-blocking. They found that areas that introduced fines for 'bed blocking' actually had higher LOS than areas that did not. The authors propose that this is due to the fines 'crowding out' the intrinsic motivation that the long-term care institutions already had for bringing down LOS.

Steventon et al. (2012) looked at two interventions requiring higher levels of primary care – intermediate care and integrated care. They compared outcomes to a control group. The intermediate care intervention aimed at supporting older people following discharge from the local general hospital. Multispecialty teams visited the wards of the hospital on a

daily basis, coordinating the discharge of patients into the care of community-based generic health workers who performed health tasks such as monitoring blood pressure and testing blood and urine. The integrated care intervention involved care management for older people. Multidisciplinary teams were established. Patients could be referred to the teams by general medical practitioners or after assessment by the local authority for support for social care needs. The study found that the intermediate care group had a higher number of unscheduled admissions into hospital and the integrated care group showed no difference. Both interventions led to higher rates of mortality than what is seen in their matched control groups.

Finally, Jacklin, Roberts et al. (2003) looked at an intervention called Virtual outreach, where a “teleconsultation” – real-time consultations where doctors and patients are separated geographically but communicate through the use of videoconferencing – are used instead of a normal outpatient visit. The authors found that over a six-month period, the costs were greater for the virtual outreach consultations than for conventional outpatient appointments – the equipment used, which might be expected to have reduced costs over time, only accounted for some of this greater cost, with increased consultant and general practitioner time accounting for the rest. There was a saving to patients in terms of costs and time. However, this did not outweigh the additional cost of the intervention.

An alternative to specialists undertaking outreach services is the concept of general practitioners with special interests, under which general practitioners specialise in a particular area and are able to run the equivalent of a specialist clinic for relevant patients. Coast et al. (2005) looked at the cost of a general practitioner with special interest service for dermatology and found that it was more costly than traditional hospital outpatient care.

2. Incentivizing primary care to reduce the use made of specialist care

Many of the papers looking at interventions that reduce hospital admissions point to ‘environmental’ factors that are deemed necessary (though not on their own sufficient) for interventions to be successful (Bodenhei-

mer et al., 2002b). In particular, any method of paying the primary care practice or practitioner will create incentives and therefore have a potential impact on behaviour. We consider the role of incentives under two headings: the implicit incentives for use of specialist care inherent in any payment mechanism, and the explicit incentives to reduce the use of specialist care embodied in what are known as ‘pay for performance’ (P4P) payment mechanisms.

2.1 Implicit incentives

Implicit incentives for GPs in relation to specialist care depend on the details of the payment scheme. For example, if a practice receives an annual capitation payment or salary to look after a registered patient population, the practice may not have a direct interest in maintaining the health of that population, and may indeed have an incentive to shift the costs of care onto other providers, such as specialists and hospitals. In the extreme, a practice might encourage patients to use emergency care inappropriately or delay intervention until hospital treatment becomes necessary. Clearly, such incentives can be moderated by the natural desire of clinicians to promote the health of their patients, and by appropriate performance monitoring and other regulatory devices. For example, the rates of emergency admissions amongst elderly patients or risk-adjusted avoidable hospital admissions might be published. To be fully effective, such instruments may require the implementation of other reforms, such as the ability of patients to easily change their GP.

The tendency for practices to encourage an ‘overuse’ of specialist care can be reversed by asking practices to manage a capitation budget for their population, from which they must purchase specialist care and other services for their patients. Under this system of ‘budget responsibility’, practices have an incentive to scrutinize the need for specialist care more critically, in order to adhere to their budget. They may then seek to delay or refuse some specialist referrals, or to treat some patients in a less costly primary care setting.

The strength of incentives under budget responsibility depends heavily on how ‘hard’ the budget constraint is made, and how much of any surplus the practice is able to retain. For example, recent attempts in England to implement ‘Practice Based Commissioning’ have largely been ineffec-

tive. One of the contributory reasons has been that in many localities, the budgets given to general practices have been notional, with few sanctions for overspending, or rewards for underspending (Audit Commission, 1996).

Between 1990 and 1998, the English NHS experimented with a scheme known as GP fundholding, under which GPs could elect to assume responsibility for a budget that covered routine non-emergency hospital treatments and pharmaceutical spending. Experiments of a similar design have been attempted elsewhere (Thorlby et al., 2011). The budgets for fundholders were quite soft, in the sense that GPs' income was not directly at risk, and surpluses had to be spent on some aspect of patient services. Deficits had few concrete adverse implications for many practices, and were often guaranteed by the paying health authority (Audit Commission, 1996).

Dusheiko et al. (2006) studied hospital use by English practices before and after the abolition of fundholding in 1998. They estimated that fundholders made 4.9 per cent less use of the relevant non-emergency hospital treatments than their non-fundholding counterparts, a difference that quickly disappeared after the abolition. A crucial question, however, is whether the reduction in utilization had any adverse impact on patients' health, an issue that researchers have been unable to definitively address. More generally, to be fully effective, any system of budget responsibility requires that health outcomes are monitored to ensure that expenditure control is not being secured at the expense of population health.

A further important issue under any system of budget responsibility for primary care is the level of risk inherent in healthcare budgets for small population groups, such as practice populations. Martin et al. (1998) estimate that under the prevailing budget-setting regime, there was a 1 in 3 chance that the annual expenditure of a typical fundholding practice (10 000 patients) would vary more than 10 per cent from the budget. Although some of this variation might be due to clinical practice or weaknesses in the budget-setting process (and therefore be amenable to improvement), some is due to the inherent stochastic nature of the need for health services, and is therefore completely beyond the control of the practice. Therefore, a very careful risk management – perhaps in the form of cost-sharing, stop-loss arrangements, or the removal of responsibility

for high cost patients – is needed to avoid exposing practices to very high levels of budgetary risk.

Other implicit incentives relevant to specialist care may arise from the nature of the market in which primary care operates. For example, within a gatekeeping system, patients may be able to choose their general practitioner. This might encourage a more parsimonious use of specialist care if the GP income depends on attracting patients, and patients perceive an unnecessary use of specialist care to be a signal of poor GP quality. On the other hand, GPs might compete on their willingness to offer patients easy access to secondary care, in which case a perversely high use of specialist care might arise. Assuming that patients choose their GP partly on basis of perceived quality, much will depend on the performance information made available to patients, and how it is presented and explained.

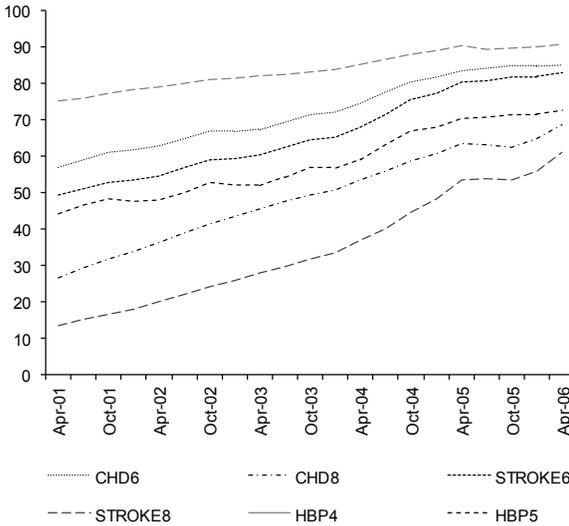
2.2 Explicit incentives

Under P4P arrangements, clinicians or practices receive a payment for meeting certain structural, process, outcomes or other performance criteria. The common feature is that the activity or result being rewarded should either be a direct measure of patient benefit (in the form of improved health or patient experience) or some action that is known (from research evidence) to lead to improved health. Measures of patient benefit might include risk-adjusted mortality rates. Close proxies for population health, such as levels of blood pressure, have also been used. Process measures, such as adherence to clinical guidelines, are in common use. The most widespread mechanisms are direct payments for specific preventive measures, such as screening, immunization and vaccination. However, health systems are increasingly seeking to reward broader indications of good healthcare delivery or health outcomes.

In 2004, the National Health Service in the United Kingdom introduced a pay-for-performance contract for family practitioners. This scheme, known as the Quality and Outcomes Framework (QOF), increased existing income according to performance with respect to about 150 quality indicators, including clinical care for 10 chronic diseases, the organization of care, and patient experience. Practices are rewarded according to the aggregate score achieved across these indicators, with up to

20 percent of practice income contingent on reported performance (Dusheiko et al., 2006). Figure 1 summarizes trends in six QOF indicators on either side of the implementation date (April 1 2004).

Figure 1. Trends in six QOF indicators 2001-2006



Source: QRESEARCH (Version 12) and the Information Centre for Health and Social Care.

Note: Key (all measured in percent of eligible population):
 CHD6 Blood Pressure < 150/90 in last 15 months for patients with CHD
 CHD8 Cholesterol < 5 mmol/l in last 15 months for patients with CHD
 STROKE6 BP < 150/90 in last 15 months for patients with stroke
 STROKE8 Cholesterol < 5 mmol/l in the last 15 months for patients with stroke
 HBP4 Blood pressure recorded in last 9 months for patients with hypertension
 HBP5 Blood pressure < 150/90 in the last 9 months for patients with hypertension

Downing et al. (2007) found that higher QOF clinical domain scores were generally associated with lower hospital admission rates. Since then, there have been several studies of QOF attainment in particular disease areas. For example, after adjusting for practice population deprivation status, Shohet et al. (2007) found that there was a significant inverse association between the proportion of epilepsy-treated seizure-free patients (calculated using a QOF indicator) and the proportion of epilepsy-treated patients with at least one epilepsy-related emergency hospitalization. For every 1 percent increase in the proportion of seizure-free epilepsy-treated patients, there was a 0.43 percent reduction in the number of patients with at least one epilepsy-related emergency hospitalization.

Dusheiko et al. (2011) found that a 10 percent improvement (judged by the practices QOF score) in the general practice quality of stroke care between 2004/2005 and 2007/2008 reduced 2007/2008 hospital expenditure by about GBP 130 million in England. The cost savings were found to mainly be due to reductions in emergency admissions and outpatient visits, rather than to lower costs for patients treated in hospital or to reductions in elective admissions. In contrast, they found little evidence of savings arising in other specialities. In the same vein, Bankart, Baker et al. (2011) found that there was no association between aggregate QOF scores and emergency admission rates.

In a paper looking at the introduction of a P4P scheme for primary care in Australia, Scott, Schurer et al. (2009) examined the impact of the scheme on the number of HbA1c tests ordered. It found that the incentive programme had a positive effect on quality of care in diabetes management. The magnitude of the effect was approximately a 20 percentage point increase in the probability of ordering an HbA1c test since the reform was introduced. The subsequent impact on hospital admissions and other use of specialist care was not reported.

Mullen et al. (2010) studied physician medical groups contracting with a large network HMO in the US to compare clinical quality before and after the implementation of two P4P programmes. The first was a quality incentive programme (QIP), which paid quarterly bonuses to medical groups performing at or above the 75th percentile from the preceding year on one or more of five clinical quality measures. One year after the QIP went into effect, the HMO joined forces with five other health plans in a coordinated P4P programme sponsored by California's Integrated Healthcare Association (IHA), a non-profit coalition of health plans, physician groups, hospitals, and purchasers. Five of the six measures selected by the IHA were also targets of the original QIP programme. The clinical service measures rewarded by the programme included cervical cancer screening, breast cancer screening, and HbA1c testing for diabetics. They did not find any association with readmissions or avoidable hospitalizations. The authors concluded that the P4P initiative did neither result in a major improvement in quality nor in a notable disruption in care. In particular, although some rewarded measures may have improved in response to the programme, they did not find any evidence of positive 'spillovers' to other aspects of care. For instance, they noted that provid-

ers that made improvements to their IT, such as automated reminders for appointments, in order to improve outcomes that were rewarded, did not appear to make the natural extension to use these IT improvements to increase performance on other measures, even when the cost was small, if there was no obvious return.

Lester et al. (2010) looked at the introduction and then the removal of a P4P programme for patients from Kaiser Permanente. The programme offered financial incentives based on a yearly assessment of patient level glycaemic control (HbA1c <8 percent), screening for diabetic retinopathy, control of hypertension (systolic blood pressure <140 mm Hg), and screening for cervical cancer. The payment attached to each indicator was directed to its relatively large medical care facilities rather than to individual doctors, and doctors' income is not affected by the incentives. The study focused on two indicators – screening for diabetic retinopathy and screening for cervical cancer. The incentives for the two types of screening were removed during the study period. During the five consecutive years when financial incentives were attached to screening for diabetic retinopathy (1999-2003), the rate rose from 84.9 percent to 88.1 percent. This was followed by four years without any incentives when the rate fell year by year to 80.5 percent. During the two initial years when financial incentives were attached to cervical cancer screening (1999-2000), the screening rate rose slightly, from 77.4 percent to 78.0 percent. During the next five years, when the financial incentives were removed, the screening rates fell year by year to 74.3 percent. The incentives were then reattached for two years (2006-2007) and the screening rates began to increase. Across the 35 facilities, the removal of incentives was associated with a decrease in performance of about 3 per cent per year on average for screening for diabetic retinopathy and about 1.6 per cent per year for cervical cancer screening. The authors concluded that removing facility directed financial incentives from clinical indicators may mean that the performance levels decline.

The academic literature has hitherto been cautious about the effectiveness of P4P programmes. Most schemes have been small scale and tentative, and researchers have struggled to find a material impact. However, there has been some success in encouraging desired behaviour. And there is a growing recognition that the provider payment mechanism is a crucial area for future experimentation. However, it is worth noting that any P4P

scheme is open to creating perverse and unintended consequences and systems are likely to be susceptible to gaming (Gravelle et al., 2009), so careful experimentation, audit and review will always be necessary.

3. Discussion and policy implications

Increases in spending on healthcare can have a positive impact on health outcomes. Where to focus additional resources remains an area with a dearth of evidence at the macro level in many countries. Martin et al. (2008) examined the link between health spending and outcomes in England in two large programmes of care (circulatory disease and cancer) and found that the marginal cost of a quality adjusted life year saved was GBP 11 960 for circulatory disease and GBP 19 070 for cancer, i.e. lower than many commentators have previously suggested.

This paper has focused on primary care interventions designed to reduce the use of specialist care, and the role of incentives in maximizing the effectiveness of such interventions. There is little persuasive evidence on the macro benefits of primary care spending, and – with a few exceptions – the micro evidence is small scale and inconclusive, although there are indications of promising policy options for future experimentation. Whatever the intervention under scrutiny, there are some further policy design issues that must be satisfied if the initiative is to stand any chance of success.

Information is a fundamental prerequisite for all successful efforts to constrain the unnecessary use of secondary care. In particular, it is essential to be able to ensure that any reduction in healthcare expenditure does not come at the expense of unacceptable quality reductions. Information on quality might be used by regulators wishing to check on GP and provider quality, on payers wishing to reward the volume and quality of outputs, by patients seeking to choose GPs and secondary care providers, by citizens wishing to hold governments and insurers to account, and by researchers wishing to evaluate cost containment initiatives. There is a fundamental coordination role for governments or their agents to devise appropriate information systems, mandate data collection, devise appropriate public reporting mechanisms, ensure data quality and tackle fraud.

More generally, high-quality governance is often an important element of any cost containment initiative. It is essential that accountability mechanisms are in place to ensure that payers, patients and citizens can take appropriate action when the performance is inadequate, either in the form of poor quality or inefficiency. Accountability might be through direct contractual relations with providers, but could also take the form of professional regulation, market mechanisms (provider choice) or local or national elections (to replace ineffective governments).

The experience in all health system reforms is that a reform stands little chance of success without clinical leadership and engagement, including at the most senior level. In the case of primary care, this implies that any reform should, wherever possible, seek an alignment with clinical preoccupations. This should be feasible with initiatives such as P4P, if the rewarded outcomes are in line with clinical values. However, it may be more challenging for initiatives such as budget responsibility. It is, for example, noteworthy that only 50 percent of the English general practitioners opted to join the (voluntary) fundholding scheme, even when it appeared to offer favourable financial rewards to GP participants.

It is important to note that disappointing results from some pilot schemes described above may be due to their small scale, or the short time for which they are implemented, or the limited range of treatments to which they apply. If either primary or secondary care providers need to make any substantial investments in order to respond to new incentives, a lack of power in the pilot studies may blunt their effectiveness.

It is also the case that many schemes have been aimed at the population at large, and that an improved focus on the subgroups of patients who stand to gain most from an intervention may secure an improved cost-effectiveness. Increasingly, it is likely that such targeting of initiatives will be a key to the optimal use of limited primary care resources.

It is intrinsically difficult to evaluate many of the potential primary care initiatives designed to secure reductions in specialist care. They will often involve an increased expenditure in the short term, the benefits of which will only become apparent in later years. Many research studies merely focus on whether a desired behavioural change in primary care has been secured, without examining the subsequent impact on secondary care utilization. Moreover, a comprehensive evaluation would not only examine the impact on specialist care, but also the broader impact of

interventions on worker productivity, welfare payments, social care and other societal costs of ill-health. Given the difficulties of empirical studies, there is a case for using microsimulation modelling (for example Spielauer, 2007 and Zucchelli et al., 2010) to examine the longer term and broader consequences of primary care interventions.

An additional concern in many countries is the fairness of distribution of healthcare resources, and inequalities in health. By improving the access to health care, an emphasis on primary care is generally considered to work in favour of poorer and sicker people (Shi et al., 1999; Baker and Middleton, 2002). Where appropriate, such equity concerns should also be built into any evaluation of new interventions.

Finally, it is worth noting that the distinction between primary and secondary care may become increasingly blurred. As the number of older people with complex chronic medical needs increases, the demand for integration of care, and personalized medical treatment will also grow. Whether there will be a provider response to such demand is likely to depend on the reform of provider payment mechanisms, particularly for secondary care. At present, these usually reward discrete episodes of care. In the future, payment mechanisms are increasingly likely to reward 'bundles' of care, or indeed a whole year of care, for people with complex needs. It will be interesting to see whether health systems can respond to these new pressures, and whether appropriate organizations will arise to manage the challenges of integrating care for such patients. Experience in the US with the new 'Accountable Care Organizations', responsible for the costs and quality of health care for a defined population (with a minimum size of 5 000 people), will be of great interest in this respect (Shortell et al., 2010).

References

- Audit Commission (1996), *What the Doctor Ordered*, The Audit Commission, London.
- Bailey, W.C., Richards, J.M., Jr., Brooks, C.M., Soong, S.J., Windsor, R.A. and Manzella, B.A. (1990), A randomized trial to improve self-management practices of adults with asthma, *Archives of Internal Medicine* 150, 1664-1668.
- Baker, D. and Middleton, E. (2002), Cervical screening and health inequality in England in the 1990s, *Journal of Epidemiology and Community Health* 57, 417-423.

- Bankart, M.J., Baker, R., Rashid, A., Habiba, M., Banerjee, J., Hsu, R., Conroy, S., Agarwal, S. and Wilson, A. (2011), Characteristics of general practices associated with emergency admission rates to hospital: A cross-sectional study, *Emergency Medical Journal* 28, 558-563.
- Bazian, L. (2005), Specialist outreach into primary care: Is it better than standard care?, *Evidence-Based Healthcare & Public Health* 9, 294-301.
- Bodenheimer, T., Wagner, E.H. and Grumbach, K. (2002a), Improving primary care for patients with chronic illness, *JAMA: Journal of the American Medical Association* 288, 1909-1914.
- Bodenheimer, T., Wagner, E.H. and Grumbach, K. (2002b), Improving primary care for patients with chronic illness – The chronic care model, part 2, *JAMA: Journal of the American Medical Association* 288, 1909-1914.
- Bolton, M., Tilley, B., Kuder, J., Reeves, T. and Schultz, L. (1991), The cost and effectiveness of an education program for adults who have asthma, *Journal of General Internal Medicine* 6, 401-407.
- de Bruin, S.R., Heijink, R., Lemmens, L.C., Struijs, J.N. and Baan, C.A. (2011), Impact of disease management programs on healthcare expenditures for patients with diabetes, depression, heart failure or chronic obstructive pulmonary disease: A systematic review of the literature, *Health Policy* 101, 105-121.
- Clark, N., Feldman, C., Evans, D., Levison, M., Wasilewski, V. and Mellins, R. (1986), The impact of health education on frequency and cost of health care use by low income children with asthma, *Journal of Allergy and Clinical Immunology* 78, 108-115.
- Cline, C.M., Israelsson, B.Y., Willenheimer, R.B., Broms, K. and Erhardt, L.R. (1998), Cost effective management programme for heart failure reduces hospitalisation, *Heart* 80, 442-446.
- Coast, J., Noble, S., Noble, A., Horrocks, S., Asim, O., Peters, T.J. and Salisbury, C. (2005), Economic evaluation of a general practitioner with special interests led dermatology service in primary care, *British Medical Journal* 331, 1444-1449.
- Domurat (1999), Diabetes managed care and clinical outcomes: The Harbor City, California Kaiser Permanente diabetes care system, *American Journal of Managed Care* 5, 1299-1307.
- Downing, A., Rudge, G., Cheng, Y., Tu, Y.K., Keen, J. and Gilthorpe, M. (2007), Do the UK government's new Quality and Outcomes Framework (QOF) scores measure primary care performance? A cross-sectional survey of routine healthcare data, *BMC Health Services Research* 7(166), 1-7.
- Dusheiko, M., Gravelle, H., Jacobs, R. and Smith, P. (2006), The effect of financial incentives on gatekeeping doctors: Evidence from a natural experiment, *Journal of Health Economics* 25, 449-478.
- Dusheiko, M., Gravelle, H., Martin, S., Rice, N. and Smith, P. (2011), Does better disease management in primary care reduce hospital costs? Evidence from English primary care, *Journal of Health Economics* 30, 919-932.
- Ellerbeck, E.F., Ahluwalia, J.S., Jolicoeur, D.G., Gladden, J. and Mosier, M.C. (2001), Direct observation of smoking cessation activities in primary care practice, *Journal of Family Practice* 50, 688-693.
- Fiore, M., Bailey, W. and Cohen, S. (2000), *Treating Tobacco Use and Dependence, Clinical Practice Guideline, U.S. D.o.H.a H. Services, Department of Health and Human Services, Rockville.*

- Godtfredsen, N.S., Vestbo, J., Osler, M. and Prescott, E. (2002), Risk of hospital admission for COPD following smoking cessation and reduction: A Danish population study, *Thorax* 57, 967-972.
- Gravelle, H., Sutton, M. and Ma, A. (2009), Doctor behaviour under pay for performance contract: Treating, cheating and case finding?, *Economic Journal* 120, F129-F156.
- Greineder, D.K., Loane, K.C. and Parks, P. (1999), A randomized controlled trial of a pediatric asthma outreach program, *Journal of Allergy and Clinical Immunology* 103, 436-440.
- Hamilton, B.A. (2007), *Key Drivers of Demand in Emergency Department*, Booz Allen Hammiton, Sydney.
- Holmas, T.H., Kjerstad, E., Luras, H. and Straume, O.R. (2010), Does monetary punishment crowd out pro-social motivation? A natural experiment on hospital length of stay, *Journal of Economic Behaviour and Organisation* 75, 261-267.
- Jacklin, P.B., Roberts, J.A., Wallace, P., Haines, A., Harrison, R., Barber, J.A., Thompson, S.G., Lewis, L., Currell, R., Parker, S., Wainwright, P. and V.O.P. Group (2003), Virtual outreach: Economic evaluation of joint teleconsultations for patients referred by their general practitioner for a specialist opinion, *British Medical Journal* 327, 84-88.
- Johansson, B., Holmberg, L., Berglund, G., Brandberg, Y., Hellbom, M., Persson, C., Glimelius, B. and Sjoden, P.O. (2001), Reduced utilisation of specialist care among elderly cancer patients: A randomised study of a primary healthcare intervention, *European Journal of Cancer* 37, 2161-2168.
- Jones, J., Wilson, A., Parker, H., Wynn, A., Jagger, C., Spiers, N. and Parker, G. (1999), Economic evaluation of hospital at home versus hospital care: Cost minimisation analysis of data from randomised controlled trial, *British Medical Journal* 319, 1547-1550.
- Kauppinen, R., Vilkkka, V., Sintonen, H., Klaukka, T. and Tukiainen, H. (2001), Long-term economic evaluation of intensive patient education during the first treatment year in newly diagnosed adult asthma, *Respiratory Medicine* 95, 56-63.
- Lahdensuo, A., Haahtela, T., Herrala, J., Kava, T., Kiviranta, K., Kuusisto, P., Pekurinen, M., Peramaki, E., Saarelainen, S., Svahn, T. and Liljas, B. (1998), Randomised comparison of cost effectiveness of guided self management and traditional treatment of asthma in Finland, *British Medical Journal* 316, 1138-1139.
- Lester, H., Schmittiel, J., Selby, J., Fireman, B., Campbell, S., Lee, J., Whippy, A. and Madvig, P. (2010), The impact of removing financial incentives from clinical quality indicators: Longitudinal analysis of four Kaiser Permanente indicators, *British Medical Journal* 340, 340:c1898.
- Lowy, A., Kohler, B. and Nicholl, J. (1994), Attendance at accident and emergency departments: Unnecessary or inappropriate?, *Journal of Public Health Medicine* 16, 134-140.
- Macinko, J., Dourado, I., Aquino, R., Bonolo Pde, F., Lima-Costa, M.F., Medina, M.G., Mota, E., de Oliveira, V.B. and Turci, M.A. (2010), Major expansion of primary care in Brazil linked to decline in unnecessary hospitalization, *Health Affairs* 29, 2149-2160.
- Macinko, J., Starfield, B. and Shi, L. (2003), The contribution of primary care systems to health outcomes within Organization for Economic Cooperation and Development (OECD) countries, 1970-1998, *Health Services Research* 38, 831-865.

- Martin, S., Rice, N. and Smith, P.C. (1998), Risk and the general practitioner budget holder, *Social Science and Medicine* 47, 1547-1554.
- Martin, S., Rice, N. and Smith, P.C. (2008), Does health care spending improve health outcomes? Evidence from English programme budgeting data, *Journal of Health Economics* 27, 826-842.
- Mattke, S., Seid, M. and Ma, S. (2007), Evidence for the effect of disease management: Is \$1 billion a year a good investment?, *American Journal of Managed Care* 13, 670-676.
- McCoy, D., Godden, S., Pollock, A.M. and Bianchessi, C. (2007), Carrot and sticks? The Community Care Act (2003) and the effect of financial incentives on delays in discharge from hospitals in England, *Journal of Public Health* 29, 281-287.
- Mullen, K.J., Frank, R.G. and Rosenthal, M.B. (2010), Can you get what you pay for? Pay-for-performance and the quality of healthcare providers, *RAND Journal of Economics* 41, 64-91.
- Munroe, W.P., Kunz, K., Dalmady-Israel, C., Potter, L. and Schonfeld, W.H. (1997), Economic evaluation of pharmacist involvement in disease management in a community pharmacy setting, *Clinical Therapeutics* 19, 113-123.
- Myers, P. (1982), Management of minor medical problems and trauma: General practice or hospital?, *Journal of the Royal Society of Medicine* 75, 879-883.
- Naidoo, B., Stevens, W. and McPherson, K. (2000), Modelling the short term consequences of smoking cessation in England on the hospitalisation rates for acute myocardial infarction and stroke, *Tobacco Control* 9, 397-400.
- National Institute for Health and Clinical Excellence (2006), National costing report – Smoking cessation in primary care, <http://publications.nice.org.uk/brief-interventions-and-referral-for-smoking-cessation-ph1>, National Institute for Health and Clinical Excellence, London.
- Oddone, E.Z., Weinberger, M., Giobbie-Hurder, A., Landsman, P. and Henderson, W. (1999), Enhanced access to primary care for patients with congestive heart failure, Veterans Affairs Cooperative Study Group on Primary Care and Hospital Readmission, *Effective Clinical Practice* 2, 201-209.
- Parchman, M.L. and Culler, S.D. (1999), Preventable hospitalizations in primary care shortage areas. An analysis of vulnerable Medicare beneficiaries, *Archives of Family Medicine* 8, 487-491.
- Patel, A., Knapp, M., Perez, I., Evans, A. and Kalra, L. (2004), Alternative strategies for stroke care: Cost-effectiveness and cost-utility analyses from a prospective randomized controlled trial, *Stroke* 35, 196-203.
- Pieber, T.R., Holler, A., Siebenhofer, A., Brunner, G.A., Semlitsch, B., Schattenberg, S., Zapotoczky, H., Rainer, W. and Krejs, G.J. (1995), Evaluation of a structured teaching and treatment programme for type 2 diabetes in general practice in a rural area of Austria, *Diabetic Medicine* 12, 349-354.
- Purshouse, R., Brennan, A., Latimer, N., Meng, Y., Rafia, R., Jackson, R. and Meier, P. (2009), Modelling to assess the effectiveness and cost-effectiveness of public health related strategies and interventions to reduce alcohol attributable harm in England using the Sheffield Alcohol Policy Model version 2.0, Report to the NICE Public Health Programme Development Group, University of Sheffield.
- Rich, M.W., Beckham, V., Wittenberg, C., Leven, C.L., Freedland, K.E. and Carney, R.M. (1995), A multidisciplinary intervention to prevent the readmission of elderly

- patients with congestive heart failure, *New England Journal of Medicine* 333, 1190-1195.
- Riegel, B., Carlson, B., Glaser, D. and Hoagland, P. (2000), Which patients with heart failure respond best to multidisciplinary disease management?, *Journal of Cardiovascular Failure*, 290-299.
- Roberts, E. and Mays, N. (1998), Can primary care and community-based models of emergency care substitute for the hospital accident and emergency (A & E) department?, *Health Policy* 44, 191-214.
- Sadur, C., Moline, N. and Costa, M. (1999), Diabetes management in a health maintenance organization: Efficacy of care management using cluster visits, *Diabetes Care* 22, 2011-2017.
- Scott, A., Schurer, S., Jensen, P.H. and Sivey, P. (2009), The effects of an incentive program on quality of care in diabetes management, *Health Economics* 18, 1091-1108.
- Shi, L., Starfield, B., Kennedy, B. and Kawachi, I. (1999), Income inequality, primary care, and health indicators, *Journal of Family Practice* 48, 275-284.
- Shohet, C., Yelloly, J., Bingham, P. and Lyratzopoulos, G. (2007), The association between the quality of epilepsy management in primary care, general practice population deprivation status and epilepsy-related emergency hospitalisations, *Seizure-European Journal of Epilepsy* 16, 351-355.
- Shortell, S.M., Casalino, L. and Fisher, E. (2010), How the Center for Medicare and Medicaid Innovation should test accountable care organizations, *Health Affairs* 29, 1293-1298.
- Spielauer, M. (2007), Dynamic microsimulation of health care demand, health care finance and the economic impact of health behaviours: Survey and review, *International Journal of Microsimulation* 1, 35-53.
- Starfield, B. (1992), *Primary Care: Concept, Evaluation, and Policy*, Oxford University Press, New York.
- Steventon, A., Bardsley, M., Billings, J., Georghiou, T. and Lewis, G. (2012), The role of matched controls in building an evidence base for hospital-avoidance schemes: A retrospective evaluation, *Health Services Research* 47, 1679-1698.
- Stewart, S., Blue, L., Walker, A., Morrison, C. and McMurray, J.J. (2002), An economic analysis of specialist heart failure nurse management in the UK. Can we afford not to implement it?, *European Heart Journal* 23, 1369-1378.
- Stewart, S., Vandenbroek, A., Pearson, S. and Horowitz, J. (1999), Prolonged beneficial effects of a home-based intervention on unplanned readmissions and mortality among patients with congestive heart failure, *Archives of Internal Medicine* 159, 257-261.
- Testa, M.A. and Simonson, D.C. (1998), Health economic benefits and quality of life during improved glycemic control in patients with type 2 diabetes mellitus: A randomized, controlled, double-blind trial, *JAMA: Journal of the American Medical Association* 280, 1490-1496.
- Thorlby, R., Rosen, R. and Smith, J. (2011), *GP commissioning: Insights from medical groups in the United States*, Nuffield Trusts.
- Vojta, C., Amaya, M. and Browngoehl, K. (1999), A home-based asthma education program in managed Medicaid, *Journal of Clinical Outcomes Management* 6(10), 30-34.

- Wagner, E.H., Austin, B.T., Davis, C., Hindmarsh, M., Schaefer, J. and Bonomi, A. (2001), Improving chronic illness care: Translating evidence into action, *Health Affairs* 20, 64-78.
- Wagner, E.H., Sandhu, N., Newton, K.M., McCulloch, D.K., Ramsey, S.D. and Grothaus, L.C. (2001), Effect of improved glycemic control on health care costs and utilization, *JAMA: Journal of the American Medical Association* 285, 182-189.
- Wilson, S.R., Scamagas, P., German, D.F., Hughes, G.W., Lulla, S., Coss, S., Char-don, L., Thomas, R.G., Starr-Schneidkraut, N., Stancavage, F.B. et al. (1993), A controlled trial of two forms of self-management education for adults with asthma, *American Journal of Medicine* 94, 564-576.
- Zucchelli, E., Jones, A. and Rice, N. (2010), The evaluation of health policies through microsimulation methods, HEDG Working Paper 10/03, University of York, York.

Comment on Beales and Smith: The role of primary health care in controlling the cost of specialist health care

Helgi Tómasson*

The *urgent search for expenditure control mechanisms* constitutes the background and topic of Beales and Smith's paper. In the paper, there is an implicit definition of the classification of health care into primary care and specialist care. The general practitioner (GP, British reference) and preventive medicine, immunizations, vaccinations, seat belts, etc., seem to be defined as primary care, and dealing with acute care, difficult diseases, accidents, hospital admission as specialist care. The paper gives a literature review of the role of primary health care. The cited literature suggests that countries with more developed primary care systems have healthier populations. Measuring the health of a population is a challenging issue. Certainly, no simple measure is obvious. The literature reports organized efforts to reduce the use of specialist care. The results of these efforts seem somewhat mixed.

Then, the authors discuss methods for incentivizing primary care. They classify the tools into implicit and explicit incentives. Among the interesting ideas of implicit incentives is the idea of having primary care institutions buy services from more specialized units. The GP fundholding experiment in England is described as an example. It seems clear that primary care institutions in this type of environment would face a risk of rare, extremely expensive patients. This would bring along the

* University of Iceland, helgito@hi.is.

need for an insurance system, so that primary care institutions could insure against catastrophes. The explicit incentive of paying for performance is popular among economists, but its implementation in the health environment is non-trivial. The big question is how to measure progress. The authors review some cases from England. The results are somewhat unclear, which is summarized in the quote “any P4P scheme is open to creating a perverse and unintended consequence ...”.

In the conclusion, the authors discuss some policy implementations. They rightly stress the point of a good administrative system, i.e., efficient bureaucracy, accounting systems, and clinical leadership and engagement. The authors have done a fine job in summarizing the fact that the results on the matter are mixed and unclear.

A possible explanation for the mixed and unclear results is, perhaps, as mentioned by the authors, that the pilot schemes are small-scale, and are implemented for a short time, etc. In my opinion, the problem is also in the data. In the era of cheap electronic data devices, there is an abundance of data of different kinds. These data do not make sense unless an appropriate statistical model capturing key features of the data generating process is designed. The data are a result of many types of sample selection processes which depend on each other in a complicated way. Therefore, the statistical model building for the observed data is extremely difficult.

An approach mentioned by the authors might be a way out. Microsimulation modelling might be a good option here. The opening line of Orcutt (1957) is, “existing models of our socio-economic system have proved to be of rather limited predictive usefulness”. Orcutt mentions difficulties in the predictions of alternative governmental actions with a long range character. In the journal, *Medical Decision Making*, Rutter et al. (2011) view microsimulation models for health outcomes. The microsimulation approach has an obvious potential here. However, in general, it is technically demanding in the sense that it requires a great deal of model-building and programming by the researchers.

Another aspect in the search for improved cost-effectiveness is the consideration of efficiency. Health care is a composite product. Therefore, the construction of an efficient frontier approach is non-trivial. Should there be one efficient frontier for primary care and another for specialist care? How should these be weighted together? It is not easy to compare a

primary care health service to a specialist care health service. Which is more efficient? An objective approach might be similar to the complicated problem described in Koop (2002). Koop sets up a Bayesian algorithm for evaluation of the efficient-frontier of a multiple output system. The example by Koop is the measurement of the efficiency of different types of baseball players. In a ball game, some are good goal scores, whereas others are good at passing the ball. Some are able to do both. It is similar in health care. The authors state that the difference between primary care and specialist care might become increasingly blurred in the future. The health care products will become more complex. Health care is team work with multiple output. Complicated measures will therefore be needed. I suggest microsimulation models, Bayesian statistical methods and formal decision theory. The literature review by the authors clearly illustrates the limits of the analysis of aggregate observational data.

References

- Koop, G. (2002), Comparing the performance of baseball players, *Journal of the American Statistical Association* 97, 710-720.
- Orcutt, G. (1957), A new type of socio-economic system, *Review of Economics and Statistics* 80, 1081-1100.
- Rutter, C., Zaslavsky, A. and Feuer, E. (2011), Dynamic microsimulation models for health outcomes: A review, *Medical Decision Making* 3, 10-18.

Payments in support of effective primary care for chronic conditions*

Randall P. Ellis** and Arlene S. Ash***

Summary

Risk adjustment models can establish appropriate payments and incentives for delivering superior primary care, particularly to people with chronic conditions, through health-based capitation and performance assessment in a patient-centered medical home (PCMH). The practical issues and administrative structures for implementing bundled PCMH payment that we discuss are relevant for single-payer Scandinavian countries as well as the US. Feasibility is supported by the “virtual all-payer” PCMH pilot of one US health plan.

Keywords: primary care, risk adjustment, patient-centered medical home, capitation, primary care activity level (PCAL).

JEL classification numbers: I13, I18, I11.

* This research was supported by The Commonwealth Fund and Verisk Health Inc. We have benefited from discussions with and input from colleagues at Verisk Health, Bruce Nash and Lisa Sasko at CDPHP, Allan Goroll and others in the Massachusetts Coalition for Primary Care Reform (MACPR), Tim Layton and Lisa Lines. The ideas expressed are our own.

** Randall P. Ellis, Department of Economics, Boston University, and Verisk Health Inc., el-lisrp@bu.edu.

*** Arlene S. Ash, Department of Quantitative Health Sciences, University of Massachusetts Medical School, and Verisk Health Inc., Arlene.Ash@umassmed.edu.

Promoting health and improving the quality of health care while controlling costs are the core objectives of every health care system. An approach that is receiving growing attention in the US seeks to change the organizational structures and incentives for primary care practitioners to enable and motivate them to do better on these dimensions. Ash and Ellis (2012) have recently described how risk adjustment can be used to dramatically change primary care payment, whereby instead of being reimbursed for each service provided (fee-for-service [FFS] reimbursement), primary care providers receive a comprehensive monthly bundled payment plus substantial performance-based bonuses. Although this payment framework could be adopted by any health care system, it is particularly well-suited for financing a practice operating as a patient-centered medical home (PCMH). We will discuss several implementation issues, describe how one health plan customized a risk-adjusted primary care capitation model to pay three practices in a “virtual all-payer” PCMH pilot, and discuss lessons for other countries.

1. Background

The American Academy of Pediatrics initiated the idea of a “medical home” in 1967 to create a central source for all medical information about a child, especially for those with special needs (Sia et al., 2004). As conceptualized by Barbara Starfield and the US Institute of Medicine, the four core functions of the medical home were to provide “accessible, comprehensive, longitudinal, and coordinated care in the context of families and community” (National Academy of Sciences, 1996). “Patient-centeredness” was added in 2001 when seven US national family medicine organizations sought to emphasize the need to manage the care of each “whole person” for whom the practice takes responsibility.

In the US, much discussion has focused on how to promote efficiency and quality by changing payment incentives for the PCMH. Goroll et al. (2007) and Goroll (2008, 2011) argue that the best payment system to support the PCMH would have: 1) a capitated budget (that is, a bundled base payment) to support all, and only, primary care activities for the practice’s panel and 2) strong performance incentives, such as potentially large bonus payments to reward practices for cost containment, clinical

quality and patient satisfaction. To function fairly and well, both the overall budget and the performance measure calculations must be risk-adjusted: the budget, so as to match each practice's resources with its patients' needs; and performance measures, so that the practice is rewarded for better-than-expected outcomes among the specific patients for whom it takes responsibility. Ash and Ellis (2012) developed risk adjustment models to support both the base and performance assessment needs of this approach.

Denmark, the Netherlands, Norway and the UK (Gosden et al., 2001) have used bundled payment to partially or fully replace FFS for primary care; however, such primary care payment has rarely been implemented in the US, particularly in the context of the PCMH. Of the 42 US PCMH pilots described on the Patient Centered Primary Care Collaborative (PCPCC) web site in June 2011 that specified payment reforms, all but one used FFS reimbursement to make most of their payments to practices (PCPCC, 2008, 2011). The most common arrangement is a small (USD 2 to USD 5) per-member per-month (PMPM) management fee add-on to FFS to support the enhanced expectations for PCMH care (Bitton et al., 2010). Performance bonus payments and rewards for achieving higher tiers of accreditation as a PCMH are common but, once more, with one exception, the words "risk adjustment" do not appear (PCPCC, 2011). Key publications by three influential medical home adopters – Kaiser Permanente (Liang, 2010), Group Health Cooperative of Puget Sound (Reid et al., 2009), and Geisinger Health Plan (Paulus et al., 2008) – also say nothing about risk adjustment.

To our knowledge, only one PCMH implementer in the US, the Capital District Physicians' Health Plan (2011) (CDPHP), has embraced risk-adjusted primary care capitation as proposed by Ash and Ellis. CDPHP is a not-for-profit network model HMO with about 350 000 members and 10 000 providers almost exclusively in New York State (Feder, 2011). Under the plan's pilot, three practices with 18 full-time-equivalent physicians initiated practice transformation in mid-2008, and payment changed to "capitation plus bonus" in January 2009. The pilot included patients with both private and public (Medicare and Medicaid) insurance. Citing estimated cost savings of USD 8 per member per month in the first two years, CDPHP recently extended the new payment system to cover 350 providers caring for over 35 000 members (Feder, 2011). Participating

practices received start-up funding to facilitate transformation, and FFS has largely been replaced by claims-based, risk-adjusted primary care capitation plus bonus payments for exceeding normative expectations. Importantly, the new payment system applies to all patients in participating practices, not just those for whom the plan accepts financial risk. CDPHP processes FFS claims for all visits, and continues to receive fees from outside insurers such as Medicare and Medicaid, but uses primary care capitation and significant bonuses to influence the provider behavior.

2. New administrative functions

The CDPHP experience plus our conversations with physicians and policy makers suggest that three functions, undertaken by one or more entities, are needed to support fundamental payment reform for the PCMH:

- Manage financial and data processing tasks. We assign this function to a medical home *administrator*.
- Act as financial guarantor. The medical home *sponsor* provides start-up money for the transition to a PCMH practice, and covers any shortfalls, thus ensuring that providers receive the funds dictated by agreed-upon formulas.
- Provide a governance structure for the sponsor and participating PCMH practices and payers that agree to the payment structure. This is a medical home *consortium*. Consortium members commit to a process for making binding decisions, for example, specifying the procedures and activities included in the medical home innovation and how performance measures and bonus payments will be calculated.

We provide more details on each function and its responsible entity below.

2.1 Role of the administrator

The administrator signs contracts with participants, pools data confidentially, calibrates and implements risk-adjustment models for base and bonus payments, shares information with stakeholders, collects survey

information, and makes risk-based payments. These functions could be performed by a regional or national health plan or payer, but will likely be more acceptable to providers and patients if conducted more locally. The administrator should be incentivized to achieve PCMH goals and authorized to enforce data sharing and data standardization while ensuring confidentiality.

In most Scandinavian countries, the government has already contracted with all physicians, so the role of the administrator could be delegated to a local authority. In the US, the administrator has the more challenging task of signing agreements with *participating practices* governing data and payment arrangements. Revenue information, for example, is needed to calculate and implement financial flows. If participation is voluntary, as in US pilots, practices may participate for diverse reasons, including: to receive start-up funds supporting practice transformation, to replace onerous FFS constraints with a steady income to support the activities that clinicians find valuable, and the opportunity to increase earnings.

In the US, the administrator must also sign agreements with *participating payers* committing them to sharing data and cost savings within the consortium. There is a potential “free-rider” problem with payers, since non-participants also benefit from their patients receiving better care. Possible inducements for joining include: the expectation that improved information can facilitate better management, the fact that only participating payers receive timely reports describing expected and actual costs and utilization, the prestige of participating in a cutting-edge reform, or the opportunity to more directly improve the incentives for plan enrollees.

2.2 Role of the sponsor

It takes time and costs money to transform a practice, so it may take a while for savings to accrue. Furthermore, information on shared savings and bonus calculations will not be available immediately. Thus, a sponsor must be prepared to support the start-up costs without any immediate financial return. In the US, the Centers for Medicare and Medicaid Services (the federal agency responsible for insuring the elderly, persons with disabilities, and the poor), or some/all of the participating payers, could function as a sponsor. During the phase-in, the sponsor ensures that

practices receive financial and technical support for activities such as training, improving medical record system functionality, new contracting, etc. Ongoing support for such activities must eventually be funded from base and bonus payments.

CDPHP was its own sponsor, bearing the cost of practice transformation for its pilot (an estimated USD 85 000 per physician), including the costs associated with treating non-CDPHP patients. TransforMed and Verisk Health Inc. assisted with implementing practice change and calculating risk-adjusted base payments, respectively (Ash and Ellis, 2012; Feder, 2011; Grumbach et al., 2009).

2.3 Role of the consortium

The consortium brings together the sponsor, participating practices and participating payers, and defines the group among which base payment funds are pooled and bonus funds allocated. Since a consortium's practices collectively generate net savings and share in it, not all practices may wish to participate in the same consortium. In a dense market area, such as a major city, several consortia might form. Greater efficiencies and geographic equity are likely if a single entity administers all consortia within fairly large geographic regions.

Consortium participants, both practices and payers, collectively determine the services included in the primary care bundle and specify how savings (losses) are shared and bonus payments determined. For institutional and data consistency, it may be easier to create consortia from existing provider networks or payer groups. However, an important goal is to achieve comparable data, with calculations and payment transfers cutting across payers and provider networks. Some consortium functions will require antitrust relief.

The consortium proposed here resembles an alternative framework that has also been recommended in the US, called an accountable care organization (ACO). Similar to a PCMH, an ACO receives a monthly capitation payment rather than fees, and takes responsibility for controlling costs and maintaining quality for a fixed panel of patients. Both entities must collect and disseminate information to participants and coordinate patient care. However, the ACO is capitated to bear financial responsibility for all medical services – inpatient, outpatient, and pharmacy –

while the bundled payment for the PCMH is only intended to cover primary care services. Strong performance payments and provider feedback to the PCMH are intended to promote the judicious use of other services, including specialty care, hospitals, diagnostic testing and pharmacy. Unlike an ACO, the PCMH consortium does not directly pay for these other services; its practices can influence their use mostly through prudent referrals, by encouraging patient self-care and, in general, by carefully managing their patients' needs.

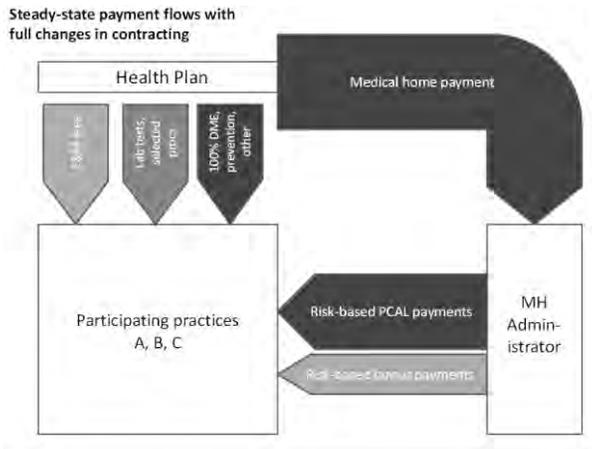
CDPHP was unusual in that it served as the sponsor, administrator, and the consortium for its PCMH pilot. As the only PCMH innovator in its primary market area, it was able to specify the scope of services to be covered by its bundled capitation payment.

3. Payment

3.1 Sample payment flows

In single payer settings (such as the Scandinavian countries), it would be relatively straightforward for the government insurance program to stop directly paying practices on a fee basis and make primary care payments directly to the Medical Home Administrator, as in Figure 1. Risk-adjusted, bundled base payments from the administrator would encourage PCMHs to creatively identify the most valuable care delivery mechanisms, including traditional office visits, group visits, emails, text messages, phone calls, and clinical and social service provision by non-physician PCMH team members. Risk-adjusted bonus payments could further encourage primary care practices to control utilization, maintain quality, and improve patient experience. Even with the payment flows shown in Figure 2, it will be important for the medical home administrator to collect enough information to enable risk adjustment and monitor performance.

Figure 1. Payment flows in a single payer system



Source: Own calculations.

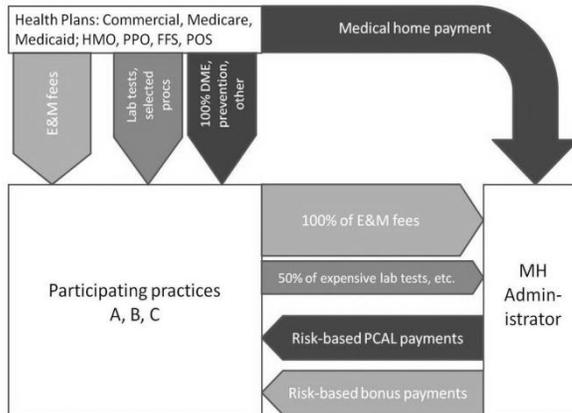
Note: E&M: evaluation and management; DME: durable medical equipment; MH: medical home; PCAL: primary care activity level.

Fees for core services and shared savings are included in the risk-adjusted bundled medical home payment, and the administrator makes base and bonus payments to the PCMH practices in place of the bulk of their FFS revenue.

Fees for core services and shared savings are included in the risk-adjusted bundled medical home payment, and the administrator makes base and bonus payments to the PCMH practices in place of the bulk of their FFS revenue.

In the US, implementing a PCMH with a new payment system is complicated by the presence of multiple payers, diverse health plans, and many complex and selective contracts between payers and providers. Although streamlined payment flows (such as in Figure 1) are desirable in the long run, Figure 2 describes a more feasible organizational structure for near-term US implementation. It requires a new organization to serve as the administrator, pooling the money and the information needed to make base payments and calculate bonuses. Figure 2 shows how payments might flow in a multi-payer medical home consortium during start-up or even longer term. Base payments would principally be financed via existing FFS payments to the PCMH, while bonus payments, to be sustainable, would eventually have to come from shared savings. Some studies suggest that PCMH savings may be achieved early (Grumbach et al., 2009; Feder, 2011).

Figure 2. Payment flows in a multi-payer, diverse benefit plan setting



Source: Own calculations.

Note: HMO: health maintenance organization; PPO: preferred provider organization; FFS: fee for service; POS: point of service; E&M: evaluation and management; DME: durable medical equipment; MH: medical home; PCAL: primary care activity level. To simplify contracting, PCMH practices continue to receive FFS payments from all payers for all services, but the revenue from evaluation and management fees and fractions of lab tests and other fees are credited by the administrator towards the PCAL base payment. Bonus payments are funded by a medical home supplement based on shared savings.

In either setting, the administrator must receive total payments that are sufficient to enable high-quality primary care. In the US, many primary care practices do not receive adequate funds that enable the more innovative forms of care (e.g., email, group meetings and expanded non-physician treatment). As calibrated in Ash and Ellis, payments need to reflect patients' expected needs, rather than the actual volume and mix of services delivered. Compared to FFS, this could change the incentives quite radically. The figures also illustrate that FFS reimbursement can be selectively used to explicitly encourage some services, such as vaccinations, by maintaining FFS billing for these services on top of the base payments. Less demonstrably useful services, including expensive primary-care-oriented laboratory tests or imaging, could be excluded from the bundled payment but only partially reimbursed at a level that covers, at most, the operating (marginal) costs. For example, the PCMH might be allowed to retain 50 percent of the full fee for certain imaging tests, with the rest having to come from its base payment. The administrator will need to monitor the spending on FFS-reimbursed services performed by

the PCMH. Lower-than-expected costs of overused services can potentially increase bonuses (CDPHP, 2011).

3.2 Payment calculations

In Ash and Ellis (2012), we provide details on how the administrator might calculate each patient's primary care activity level (PCAL), so as to ensure that practices receive the sum of their patients' PCALs. The payment for each person for each eligible month is the product of a PCAL normalized risk score (nRS) – reflecting the relative resources needed by each patient based on their age, gender, and diagnoses and expressed as a fraction (multiplier) of the average resources needed – and the average PMPM cost for delivering high-quality primary care.

The average PMPM could be calculated by dividing the total available dollars for base payments by the number of member-months covered. For example, if the agreed-upon PCMH spending pool were USD 20 million for 500 000 member months, the base payment for a patient with PCAL = 1 would be $(\text{USD } 20 \text{ million} / 500\,000) = \text{USD } 40 \text{ PMPM}$. This base amount can also be adjusted to reflect benefit plan and payer pricing differences, as further discussed below.

3.3 Accommodating diverse payers

One major hurdle to implementing primary care capitation in the US is incorporating payers and health plans with diverse benefit features and fees for the same services. For example, the allowed charges for Medicare and Medicaid patients are typically below those for commercial patients, and health maintenance organizations (HMOs) often negotiate discounts for various procedures. Furthermore, payers differ in their benefit coverage: some pay a fixed percentage of the allowed fee, while others require deductibles and fixed fees. How can bundled payments be implemented in the face of payer diversity?

For its PCMH pilot, we helped CDPHP modify the Verisk Health PCAL risk score to recognize differences in its revenues among federal and state government and employer premiums. Rather than choosing a single multiplier B for all patients, the CDPHP multiplier varied by payer category (see Table 1). Informed by its own internal regressions and other

calculations based on benefit design, CDPHP calculated the initial base payments as $A+B*nRS$, where A and B are as shown in Table 1. Even though all pilot practices achieved Level 3 accreditation as a PCMH from the US National Center for Quality Assurance (NCQA) and implemented state-of-the-art electronic health records, CDPHP continued to use claims-based diagnoses, rather than electronic health records data, mainly because of the inadequacies of the electronic systems (Feder, 2011). Future implementers may wish to investigate further refinements to the payment formula to better account for cost sharing in PCMH base payments.

Table 1. Formulas used by the CDPHP to transform a PCAL normalized risk score into a payer-specific base payment

	Minimum (A)	PCAL nRS slope (B)
Commercial HMO	\$128.80	\$60.69
Commercial non-HMO	\$105.16	\$49.65
Medicare	\$101.83	\$48.08
Medicaid	\$90.74	\$42.74

Example:
 A Medicaid patient with a PCAL of 1.8 would generate an annual base payment of $\$90.74 + (1.8 * \$42.74) = \$167.67$ (or $\$13.97/\text{month}$)

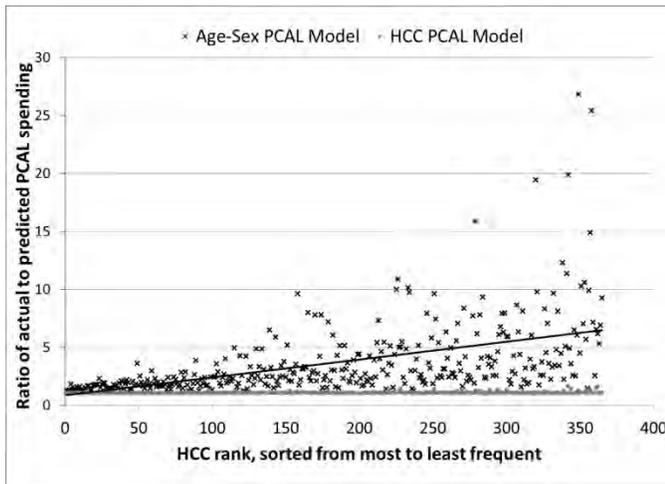
Source: Own calculations.

4. The importance of rich risk adjustment models

As highlighted in the introduction, few PCMH pilots in the US are linked to substantial payment reforms. Moreover, with the exception of CDPHP, only simple models (based on at most age-sex and the presence/absence of a few chronic conditions) are being used to adjust modest per-patient-per-month supplementary payments to FFS. When bundled payments are large, a weak risk adjustment creates a strong incentive for practices to avoid individual patients expected to cost more than the bundled payment. The PCAL payment approach predicts the primary care resources needed using binary flags that signal the presence or absence of 394 medical conditions, called Hierarchical Condition Categories or HCCs for each patient. The US government uses an HCC modeling framework to

calculate payments to private Medicare Advantage plans for their elderly and disabled Medicare enrollees; in Germany, a similar Hierarchical Morbidity Group (HMG) calculation is used to allocate health care money across its sickness funds. Figure 3 shows the ratio of actual to expected spending for each of the 365 condition categories with more than 500 cases (among 17.4 million commercially insured individuals in this sample). The widely scattered X's show the performance of a model using only age and sex to predict PCAL, while the dots hovering around 1 are for the PCAL model using HCCs, age and sex. Actual PCAL expenditures are about 50 percent higher than the age-sex predictions for the most common conditions and are progressively less accurate for rarer HCCs, while the HCC model's payments, by design, are about right for all HCCs. Figure 4 replicates the analysis using the Medicare program's 70 HCCs to predict our PCAL; while this model does better than age and sex alone, it systematically underpays for many conditions, potentially penalizing practices that care for people with these medical problems.

Figure 3. Ratio of actual to predicted PCAL for those 365 condition categories with more than 500 persons each, using the age-sex and HCC models to predict the PCAL proxy

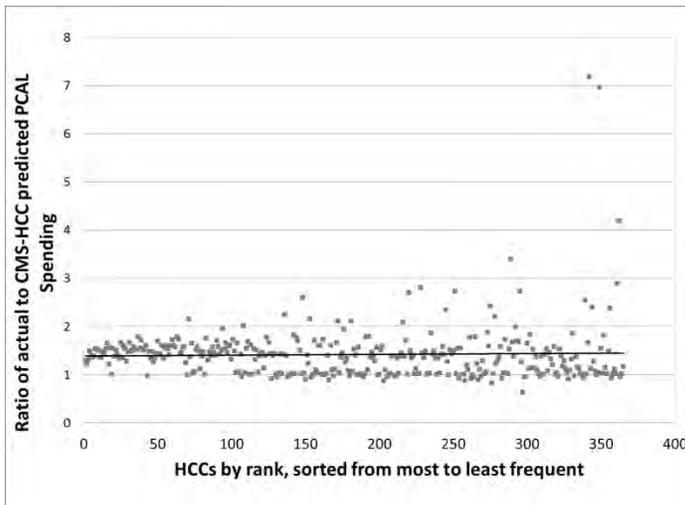


Source: Own calculations.

Note: PCAL: primary care activity level; HCC: hierarchical condition category. Regression models predicting the PCAL proxy variable (Y) were estimated using the full sample of 17.4 million people: 1) using only age and sex and 2) using age, sex, and HCCs. Per capita averages were calculated for each model for each of the 394 HCCs, based on actual and predicted PCAL costs for all people with at least one diagnosis in that HCC. HCCs were sorted from most common to least common; each data point is the ratio of actual to predicted spending, shown for 365 HCCs with more than 500 cases, ranging from HCC383 = Screening/ Observation/Special Exams with 750 471 people at the far left to HCC213 = Heart Transplant Complications with 561 people, at the far right.

In Ash and Ellis (2012), we show that customized risk adjustment is also important for assessing performance. First, using only age and sex as predictors leaves a great deal of potentially predictable variation unexplained, creating incentives for practices to avoid taking responsibility for treating the most challenging, chronically ill patients. Second, many performance outcomes (such as total spending, emergency department use and measures of patient satisfaction) may be only weakly correlated with each other, and may have different predictors. Thus, when seeking to hold a practice accountable for the difference between “what would be expected for a particular outcome with a particular patient panel” and the panel’s actual outcome, an outcome-specific regression model should be used to determine what would be expected.

Figure 4. Ratio of actual to CMS-HCC predicted PCAL for those 365 condition categories with more than 500 persons each, using the CMS-HCC model to predict the PCAL proxy



Source: Own calculations

Note: PCAL: primary care activity level; HCC: hierarchical condition category. Regression models predicting the PCAL proxy variable (Y) were estimated using the full sample of 17.4 million people using the 70 CMS-HCCs with 22 age and sex dummy variables. Per capita averages of actual and predicted PCAL were calculated for each of the 394 HCCs, for all people with at least one diagnosis in that HCC. HCCs were sorted from most common to least common; each data point plots the ratio of actual to predicted spending, shown for 365 HCCs with more than 500 cases, ranging from HCC383 = Screening/ Observation/Special Exams with 750 471 people at the far left to HCC213 = Heart Transplant Complications with 561 people, at the far right.

Table 2 illustrates how widely PCAL payments might vary for patients with differing levels of chronic or acute conditions. For example,

while a very healthy person may only require a small fraction of the average level of resources, the PCAL model predicts that a 56-year-old male with uncomplicated diabetes and fluid and electrolyte disorders can be expected to use more than three times the average level. Further, a patient with ophthalmic manifestations from diabetes, back pain and high cholesterol might require more than five times the average level of primary care. Clinically-detailed risk adjustment is needed to capture such variations and reward the PCMH for taking on these complex patients.

Table 2. PCAL payment examples for four patients

Male, Age 16	PCAL nRS = .135	Annual payment = \$65
<ul style="list-style-type: none"> • No medical problems 		
Female, Age 11	PCAL nRS = .557	Annual payment = \$267
<ul style="list-style-type: none"> • Other Non-Chronic Ear, Nose, Throat, and Mouth Disorders • Other Dermatological Disorders 		
Male, Age 56	PCAL nRS = 3.061	Annual Payment = \$1,469
<ul style="list-style-type: none"> • Benign Digestive or Urinary Neoplasm • Diabetes with no complication • Fluid/Electrolyte/Acid-Base Imbalance • Ulcer with Perforation/Obstruction • History of Disease 		
Female, Age 50	PCAL nRS = 4.791	Annual Payment = \$2,300
<ul style="list-style-type: none"> • Diabetes with Ophthalmologic Manifestation • Hyperlipidemia • Endocrine/Metabolic Disorder • Lower Back Pain • Pelvic/Uterine Inflammation • Rehab • Screening • Surgical Misadventure or Complication 		

Source: Own calculations.

4.1 Short-, medium-, and long-term needs for data

Although the risk assessments that provide information for base and bonus payments should ideally use rich data, PCMH implementers must walk before they can run; CDPHP initially chose to use only insurance claims information and rely on continuing FFS billing. Substantial continuity in data requirements is especially helpful during the start-up, since contracting arrangements, patient assignments, and practice behaviors are already in flux.

Later, as bills become unlinked from payments, we must guard against losing the key information currently found in FFS billing, specifically medical diagnoses and procedures coded in standardized formats. This requires developing a standard for “encounter records” or dummy bills and providing incentives to ensure their quality. It is also important to capture key clinical outcomes – such as blood pressure and lipid levels – currently buried in non-standardized, poorly configured electronic medical record systems and measures of the patients’ experience of care (such as is collected in the Consumer Assessment of Healthcare Providers and Systems [CAHPS] survey) of the Agency for Healthcare and Quality (AHRQ, 2011). CDPHP was using the CAHPS survey even prior to the PCMH pilot.

Eventually, it will be highly desirable to:

- Incorporate additional risk factors (such as socioeconomic status, housing or food instability, language and literacy barriers, and more detailed information on prior health status);
- Learn how to identify problems with data capture and, potentially, fraud; and
- Measure, track, and help establish the value of new kinds of health care utilization (including email and phone contacts, behavioral health interventions, and health coaching) and learn how to provide feedback to multiple stakeholders.

4.2 Coding creep

A concern with linking payments to diagnosis-based risk adjustment is that practices can increase their payments through aggressive coding. While auditing to detect fraud should be part of any health care payment system, even without fraud, payment-driven increases in coding intensity are well documented (Rosenberg et al., 2000; Angeles and Park, 2009). In response, we note that the map from diagnostic codes to measured illness in these models was designed to limit the sensitivity to variations in coding. Further, where such maps are used in the Medicare Advantage program, the government regularly recalibrates payment formulas to undo the extra money that program-wide increased coding intensity would otherwise entail. With recalibration, only *differential* upcoding under-

mines the purpose of the risk adjustment, which provides each practice with the resources its patients need. As US states assemble all-payer databases and data from patient medical records become standardized, it should be easier to achieve comparability across practices. Finally, patient care can benefit from increased attention to diagnosing and tracking medical conditions. At the same time, the incentives to over-code illnesses seem less harmful than the FFS incentives to over-provide low-value, but well-reimbursed, services.

4.3 Retrospective patient assignment and reconciliation

A fundamental challenge to implementing an all-payer PCMH in the US is that many health plans, including conventional Medicare, do not require enrollees to designate a single practice as their primary care practitioner (PCP). Another concern is that a PCMH might game the payment system by enrolling patients (to receive their bundled payments), but then focus their efforts and billing on patients not assigned to their PCMH (for whom they could continue to bill). How can we start a PCMH in the US even in the current, highly imperfect setting in which patients are not required to choose a PCP as their sole source for primary care and some (but not all) PCPs are transforming to a PCMH?

CDPHP chose not to require that patients definitively pick a PCMH practice; it used an *ex post* (retrospective) assignment algorithm that assigns patients to the practice that provided the plurality of their primary care. A similar *ex post* approach has been used in other studies, some of which have found that the patient choice of PCPs and the performance measures of these PCPs are relatively stable across years (Medicare Payment Advisory Commission, 2009; Ginsburg, 2011). Less stability has been reported by Mehrotra et al. (2010) when they assigned individual episodes to specific physicians under 12 attribution rules. While a variety of assignments are possible, what seems most attractive for the PCMH is that patients be assigned to a participating PCMH at the end of a year based on the plurality of their qualifying primary care visits or other contacts such as emails, telephone calls, or home visits (Sorbero et al., 2006). For a patient with no provider contact in a given year, the norm is to use information from the previous year. Patients with no contact with a PCMH over a two-year period presumably generated little to no primary-

care-related activities and remain unassigned. Allowing patients to select practices either within or outside the PCMH consortium facilitates the implementation, particularly in multi-payer settings where most plans do not require primary care provider selection. It is plausible that switching practices could be infrequent and as manageable with *ex post* as with *ex ante* assignment. Risk adjustment seeks to mitigate the selection problem by making payments match the patients; the goal is to make providers *financially* indifferent to whether their patients are sicker or healthier.

Ex post assignment also defeats gaming by “patient swapping,” where two practices each receive base payments for their previously assigned patients, but also earn FFS payments when they each provide most of the care for patients who “belong to” the other practice. With *ex post* assignment, regardless of which practice a patient was considered to belong to, the practice that submits bills for most of her FFS primary care services will get all of her bundled payment. *Ex post* patient assignment also mitigates the incentive to stint on providing care: if a practice underserves an enrolled patient, then either the patient 1) remains assigned to that practice, which receives her base payment but also potentially a poor service rating that worsens its bonus measures, or 2) switches to another PCP, causing her bundled payments to be redirected there. It is not clear whether the generosity of the bundled PCP payment will induce competition that results in too many or too few primary care services being provided. However, the incentives will differ from traditional FFS payment, and these bundled prices give payers a new pricing tool for promoting efficiency.

Although a prospective framework is possible, to ensure that payments more closely match the needs of practices (which should not be placed at too much financial risk), we have proposed concurrent risk adjustment for base and performance measures. In addition, changes in plan enrollment and patient assignment to practices cannot be prospectively determined. Thus, payments will need to be made based on preliminary estimates and reconciled later. However, retrospective reconciliation is a necessary feature of any bonus payment system, and is already being used for bundled payment in the Medicare Advantage and CDPHP programs. It requires no fundamental change in payment practice.

5. Conclusion

We have described an administrative structure to support a patient-centered medical home (PCMH) in either the single-payer systems in Scandinavian countries or the US multipayer setting with diverse health plans and contractual arrangements. Central to our discussion are three key administrative functions: sponsorship, information processing, and collaborative contracting, assigned, respectively, to a *sponsor*, a *medical home administrator*, and a *medical home consortium*. These entities could be existing or new organizations, depending on the organizational context.

The goals of primary care payment reform are to improve health and health care and reduce costs. Early results from PCMH pilots suggest that practice transformation is feasible and may be able to achieve these goals (Grumbach et al., 2009). To promote the adoption of bundled payments, sponsors may initially guarantee that total practice base payments are at least as large as existing practice revenue, but eventually base payments should go up or down to reflect the “effective size” (that is, the number and the complexity) of the practice. Given their increased responsibilities, bonus plus base payments seem likely to require greater payments to primary care practices than they currently receive – which currently are only about 6 to 7 percent of total health care spending in the US (Arvanites, 2009). The hope is that increasing primary care spending by perhaps as much as 2 percent of total spending can save more than that through avoidable emergency room visits, imaging, tests, and hospitalizations. Given that PCPs in Scandinavia receive a higher proportion of total health resources than in the US, less of an increase in PCP payments may be needed; however, the change from FFS to bundled payment may still create powerful new incentives.

This paper provides many specifics on how payment reform for the PCMH might be implemented. We take heart in this daunting transformation, in how one early adopter was able to deal with the complexities and thrive accommodating existing price discounts and cost sharing of diverse payers and insurance plan types with different contractual fees received. While we do not claim that the implementation we describe is in any sense optimal, it is demonstrably feasible. This detailed description of how transformative the payment for primary care can be and how it has

been implemented shows one way in which the PCMH ideal of cost-effective, high-quality primary care can emerge in the absence of centrally-dictated payment reforms.

References

- Agency for Healthcare Research and Quality (2011), Consumer Assessment of Healthcare Providers and Systems (CAHPS), <https://www.cahps.ahrq.gov/default.asp> (June 6, 2011).
- Angeles, J. and Park, E. (2009), Upcoding problem exacerbates overpayments to Medicare Advantage Plans – Administration action and house health reform bill seek to address problem, Center on Budget and Policy Priorities, <http://www.cbpp.org/files/3-12-09health.pdf> (January 23, 2011).
- Arvantes, J. (2009), IBM plans to cover its employees' deductibles, copays for primary care services, American Academy of Family Physicians News Now (November 2009), <http://www.aafp.org/online/en/home/publications/news/news-now/professional-issues/20091117ibm-prim-care.html> (June 21, 2011).
- Ash, A.S and Ellis, R.P. (2012), Risk-adjusted payment and performance assessment for primary care, *Medical Care* 50, 643-653.
- Bitton A., Martin, C. and Landon, B. (2010), A nationwide survey of patient centered medical home demonstration projects, *Journal of General Internal Medicine* 25, 584-592.
- Capital District Physicians' Health Plan (CDPHP) Enhanced Primary Care (2011), <http://www.cdphp.com/Providers/Programs/Enhanced-Primary-Care> (May 3, 2012).
- Feder, J.L. (2011), A health plan spurs transformation of primary care practices into better-paid medical homes, *Health Affairs* 30, 397-399.
- Ginsburg, P.B. (2011), Spending to save – ACOs and the medicare shared savings program, *New England Journal of Medicine* 364, 2085-2086.
- Goroll, A.H. (2011), Payment reform to support lasting practice reform in primary care, *Journal of Ambulatory Care Management* 34, 33-37.
- Goroll, A.H. (2008), Reforming physician payment, *New England Journal of Medicine* 359, 2087-2090.
- Goroll, A.H., Berenson, R.A., Schoenbaum, S.C. and Gardner, L.B. (2007), Fundamental reform of payment for adult primary care: Comprehensive payment for comprehensive care, *Journal of General Internal Medicine* 22, 410-415.
- Gosden, T., Forland, F., Kristiansen, I.S., Sutton, M., Leese B., Giuffrida, A. et al. (2001), Impact of payment method on behaviour of primary care physicians: A systematic review, *Journal of Health Services Research and Policy* 6, 44-55.
- Grumbach, K., Bodenheimer, T. and Grundy, P. (2009), Outcomes of implementing patient centered medical home interventions: A review of the evidence on quality, access and costs from recent prospective evaluation studies, http://www.pcpc.net/files/evidenceWEB%20FINAL%2010.16.09_1.pdf (December 16, 2010).
- Liang, L.L. (2010), *Connected for Health: Using Electronic Health Records to Transform Care Delivery*, Jossey-Bass, San Francisco, CA.

- Medicare Payment Advisory Commission (2009), Report to the Congress: Medicare payment policy: http://www.medpac.gov/documents/Mar09_entireReport.pdf (June 6, 2011).
- Mehrotra, A., Adams, J.L., Thomas, J.W. and McGlynn, E.A. (2010), The effect of different attribution rules on individual physician cost profiles, *Annals of Internal Medicine* 152, 649-654.
- National Academy of Sciences (1996), Primary care: America's health in a new era, National Committee for Quality Assurance (n.d.), Washington, DC, NCQA Patient-Centered Medical Home, <http://www.ncqa.org/Portals/0/PCMH%20brochure-web.pdf> (January 31, 2012, from NCQA).
- Patient-Centered Primary Care Collaborative (2008), Patient-centered medical home building evidence and momentum: A compilation of PCMH pilot and demonstration projects, Washington, DC, http://www.pcpcc.net/content/pcpcc_pilot_report.pdf (June 6, 2011).
- Patient Centered Primary Care Collaborative (2011), Pilots & demonstrations (self-reported), Washington, DC, <http://www.pcpcc.net/pcpcc-pilot-projects> (June 6, 2011).
- Paulus, R.A., Davis, K. and Steele, G.D. (2008), Continuous innovation in health care: Implications of the Geisinger experience, *Health Affairs* 27, 1235-1245.
- Reid, R., Fishman, P., Yu O. et al. (2009), Patient-centered medical home demonstration: A prospective, quasi-experimental, before and after evaluation, *American Journal of Managed Care* 15, 71-87.
- Rosenberg, M.A., Fryback, D.G. and Katz, D.A. (2000), A statistical model to detect DRG upcoding, *Health Services and Outcomes Research Methodology* 1, 233-252.
- Sia, C., Tonniges, T.F., Osterhus, E. and Taba, S. (2004), History of the medical home concept, *Pediatrics* 113, 1473-1478.
- Sorbero, M.E., Damberg, C.L., Shaw, R. et al. (2006), Assessment of pay-for-performance options for medicare physician services, Final report, RAND Corporation, Santa Monica, CA, http://www.rand.org/pubs/working_papers/2010/RAND_WR391.sum.pdf.

Comment on Ellis and Ash: Payment in support of effective primary care for chronic conditions

Jørgen T. Lauridsen^{*}

Commonly, the responsibilities and financing of health treatment are extremely fragmented. This is the essential point of departure of the study, and it is observed that such fragmentation leads to many undesirable shortcomings – inefficiency in the utilization of resources, quality loss, increased risks for patients and adverse effects just to mention a few. Per definition, this is the case in the US health care system, but it also holds true for Scandinavian countries due to a traditional practice of decentralization of many decisions regarding health care delivery as well as its financing. For standard (i.e. uniquely coded) diagnosis treatment, there is a long tradition for overcoming the aforementioned shortcomings during the implementation of concepts like shared care, integrated care etc. However, for non-standard diagnoses (i.e. diagnoses that cannot be uniquely coded), practice lags considerably behind. This does, in particular, cover many chronic conditions – type II diabetes just to mention one well-known example.

The present study aims at filling this gap by offering a three-legged framework for the improvement of responsibility handling and financial management in order to ensure three tasks: A start-up transformation of the fragmented system for finance and management (including the payment for start-up); the integration of financial and data processing tasks

^{*} Centre of Health Economics Research (COHERE), University of Southern Denmark, jtl@sam.sdu.dk.

management; and the continual governance and development of these two and related tasks.

Generally, given the lagging-behind in the development of continuous care for non-standard treatments of chronic diseases in particular, this study appears highly relevant for economic policy and decision makers. Furthermore, the suggested operational plan seems well described, sufficiently detailed and convincing. Especially, the study seems to be well founded in existing developments, including several contributions from the authors; see, for example, Ash and Ellis (2012) and further references therein.

A couple of suggestions for future developments or extensions to the study follow. First, while the proposal is carefully argued and appears convincing, it is still cast in soft terms. Specifically, it may win from simulation studies aiming at analyzing its performance under varying assumptions and circumstances. Second, given that the study is naturally embedded in a (macro) economic policy setting, some system considerations may relevantly be considered. In particular, this could involve the trade-off effect between investments in improved treatment and investments in onset prevention. For a stimulating study with an introduction to such approaches and further references, see Homer and Hirsch (2006).

References

- Ash, A.S. and Ellis, R.P. (2012), Risk-adjusted payment and performance assessment for primary care, forthcoming in *Medical Care*.
- Homer, J.B. and Hirsch, G.B. (2006), System dynamic modeling for public health: Background and opportunities, *American Journal of Public Health* 96, 452-458.

An economic assessment of price rationing versus non-price rationing of health care^{*}

Luigi Siciliani^{**}

Summary

In the presence of excess demand, health care has to be rationed in one way or another. This study reviews the relative merits of three different forms of rationing: i) price rationing, which takes the form of a copayment or a coinsurance rate; ii) rationing by waiting, when a patient is placed on a waiting list before receiving treatment; and iii) explicit rationing, when the patient is explicitly refused treatment. Both waiting times and copayments can help contain excess demand, though the demand is generally inelastic with respect to waiting times and copayments (elasticities of -0.1 or -0.2). Explicit rationing can potentially generate a higher patients' welfare as compared to copayments or waiting times, but its implementation faces several challenges.

Keywords: copayments; waiting times; explicit rationing.

JEL classification numbers: I12.

^{*} I would like to thank Mickael Bech, Sverre Kittelsen, Tor Iversen, Terkel Christiansen and participants at the NEPR conference on Economics of Health Care "Challenges in Health Care Financing and Provision" held in Reykjavik on May 7 2012 for helpful comments and suggestions.

^{**} Department of Economics and Related Studies, University of York, luigi.siciliani@york.ac.uk.

In the presence of (public or private) insurance and excess demand, health care has to be rationed in one way or another. Excess demand is likely to grow in light of reduced budgets driven by economic downturn, an ageing population and the development of new medical technologies.

In this study, I review the relative merits of three different forms of rationing in the health sector: i) price rationing, which takes the form of a copayment or a coinsurance rate; ii) rationing by waiting, when a patient is placed on a waiting list before receiving treatment; and iii) explicit rationing, when the patient is explicitly refused treatment. Both waiting times and explicit rationing are forms of non-price rationing.¹

Section 1 discusses the role of copayments or coinsurance rates within an insurance framework (ex post moral hazard about the consumption of health care). Section 2 introduces waiting times as an alternative mechanism to deal with excess demand and compares them with copayments. Section 3 argues that in general, explicit rationing can increase patients' welfare compared to the two other forms of rationing but there are several factors that impede its implementation in practice. Section 4 summarises and discusses the key findings.

1. Price rationing

In the presence of (public or private) insurance and limited supply, excess demand is likely to arise in many health systems. Economists are often quick to suggest copayments and coinsurance rates as the optimal way of dealing with such excess demand.

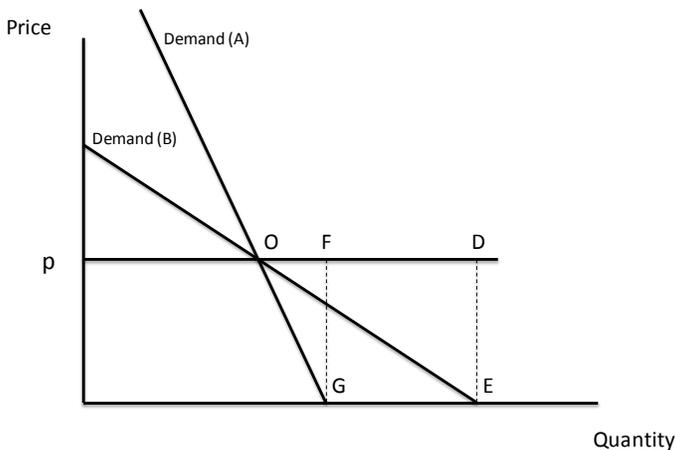
The suggestion arises from applying standard results from the theory of optimal insurance under “ex post moral hazard”, where ‘ex post’ refers to consumption ‘after’ the illness has occurred, and ‘moral hazard’ refers to the incentive to over-consume health services when prices are removed (Pauly, 1968; Zeckhauser, 1970; Zweifel et al., 2009; McGuire, 2011).

¹ For an introduction to rationing in the health sector, see Siciliani (2012). Reviews of the literature on rationing by waiting can be found in Iversen and Siciliani (2011) and Siciliani and Iversen (2012). A detailed discussion of price rationing can be found in McGuire (2011).

1.1 Optimal copayments and demand elasticity

Insurance theory provides clear insights into how optimal copayments should be set: optimal copayments should be larger the more elastic is the demand for health care. The intuition is simple: when demand is elastic, the deadweight loss associated with overconsumption is larger and a copayment (or a coinsurance rate) can be an effective way of reducing it. Instead, if demand is inelastic, then demand for health care does not respond to prices and therefore the scope for reducing overconsumption is limited. This is illustrated in Figure 1. Demand curve B is more elastic than demand curve A, and the corresponding deadweight loss (ODE) is higher than (OFG). Critically, copayments have to be traded off with consumers' losses from lower insurance levels against the risk of payment. Otherwise, it would be optimal to set the price equal to the full marginal cost. Therefore, risk aversion also plays a key role, with higher aversion reducing the scope for higher copayments. More technically, the optimal copayment is lower the higher is the difference between the marginal utility of other consumption (excluding health care) in the healthy state and the sick state.

Figure 1. Different demand elasticities



Source: Own graph.

There are (at least) two ways of making use of the normative results arising from optimal insurance theory. The first is to qualitatively differentiate copayments across different types of care on the basis of presumed

or measured elasticity. For example, it is plausible that demand is inelastic for certain types of care (for example surgery or inpatient care) and more elastic for other types of care (specialist visits, drugs and dental care) and copayments should therefore be smaller for the first type and higher for the second type. To some extent, this is consistent with what we observe. Inpatient care is typically characterised by low copayments, while the copayments for drugs or dental care can be substantial. The second possibility is to directly compute the optimal copayments as a function of the parameters of the insurance model. This is more challenging since it requires specifying the utility function of the individuals and estimates of risk aversion.

An extensive empirical literature has estimated the elasticity of demand in the health sector. The Rand experiment in the US is one of the most well-known studies in this area. It provides demand estimates by randomising individuals in different insurance plans. Therefore, it eliminates potential endogeneity biases due to individuals selecting different types of insurance packages: the concern is that sicker patients will select insurance contracts with lower coinsurance rates. Moreover, the coinsurance rates vary substantially, taking the values of zero, 25 per cent, 50 per cent and 95 per cent (Manning, Newhouse et al., 1987). The study suggests that, as compared to an individual who pays 95 per cent of health care, an individual with zero coinsurance rate has a 67 per cent higher number of doctor's visits and outpatient spending, 30 per cent higher inpatient spending, 80 per cent higher spending for dental care visits and 45 per cent higher total spending. Overall, the study suggests that demand is inelastic with an overall elasticity of around -0.1 or -0.2 . The work by Manning and Marquis (1996) combines such elasticities with estimates of risk aversion and suggests that according to optimal insurance theory, the coinsurance rates can be as high as 50 per cent, which is higher than what we observe in practice (see also Feldman and Dowd, 1991).

Standard insurance theory assumes that different types of health care, and their demands, are independent of each other so that the demand for surgery is not affected by the consumption of drugs. Goldman and Philipson (2007) argue that in some cases, the demand for different types of care can either be substitutes or complements and therefore, this has to be taken into account in the design of optimal copayments. Since the demand for drugs tends to be elastic, standard theory suggests that copay-

ments should be high. However, if drugs reduce the probability of inpatient (expensive) care, physician visits or emergency visits, it may be optimal to keep the copayment for drugs smaller. Some empirical evidence supports that such substitution between types of care is relevant (though the results are somewhat mixed; Goldman and Philipson, 2007).²

Nyman (1999a, 1999b, 2012) suggests that standard estimates of moral hazard are overstated. This is because the deadweight loss from moral hazard is computed by considering as the benchmark case what the patient would have bought without insurance. This is the wrong comparison since it includes both the *income transfer* effect due to insurance (which transfers income across health states) and the *price* effect which also arises from insurance coverage (which reduces the price paid by the consumer). The moral hazard effect should only include the price effect (excluding the income effect due to insurance). The relevant comparison is how much would the patient have chosen with insurance and health care being contractible (which cannot be empirically observed).³

1.2 Equity concerns

In publicly funded systems, the scope for using copayments is limited for equity reasons. Public health insurance systems tend to be highly redistributive. Two key principles are that health care provision should be based on need, and health care funding should be based on the ability to pay, with richer people paying more. The redistributive component is further reinforced by the fact that poorer people on average have a higher probability of falling ill (Cremer and Pestieau, 1996). The extensive use of copayments risks compromising the first principle. Indeed, possibly the

² There may also be inter-temporal substitution or complementarity effects generated by copayments. For example, an increase in copayments may reduce consumption today at the expense of higher consumption in the future. The effect may be larger for time-inconsistent individuals which could potentially translate into an overall increase in consumption.

³ As an example, consider a patient who has breast cancer and needs a mastectomy which costs USD 20 000. Suppose that without insurance the patient spends this amount. With insurance (paying a premium of USD 4 000 per year) she is fully covered and also has breast reconstruction (another USD 20 000) plus some extra days in hospital (USD 4 000). The total cost is USD 44 000. How large is moral hazard? We may be tempted to answer that it is USD 44 000-USD 20 000 = USD 24 000? According to Nyman, what we should ask is how much she would have spent if she could buy insurance and specify the amount of medical care. Suppose that the insurer gives her USD 44 000 when sick, would she spend it all or would she spend less? Suppose that in that case she would buy mastectomy and breast reconstruction: USD 40 000. Then, the moral hazard is now much smaller: USD 44 000-USD 40 000 = USD 4 000 (Nyman, 2012).

most common argument against copayments is that poor people will be more deterred from demanding treatment than the rich, thus exacerbating disparities in the receipt of health care. The Rand experiment does indeed confirm that richer individuals consume more, especially when the coinsurance rate is higher (Manning, Newhouse et al., 1987). For example, as compared to individuals in the highest third income group, individuals in the lowest third income group spend 2 per cent less if care is free and 9 per cent less if the coinsurance rate is 25 per cent. Similarly, the probability of any use is higher for the third richest income group than for the third poorest income group, respectively, equal to 90.1 per cent and 82.8 per cent with a gap of 7.3 per cent. The gap increases to 12 per cent when the coinsurance rate is 25 per cent. The result that richer (and more educated) individuals consume more health care has been found in several European studies even in the absence of (or low) copayments (Wagstaff and van Doorslaer, 2000).⁴ This is to some extent surprising given that richer individuals generally have better health. It also suggests that a higher copayments risk reinforces disparities in health utilisation across income groups.

One way of minimising potential disparities generated by copayments is to design income-tested copayments or exempt the more vulnerable groups (as is done in some countries). There may, however, be administration costs involved in pursuing this. Moreover, the eligibility criteria need to be carefully designed and easily enforced. If income-tested, the declared levels of income need to be reliable and verifiable. Uniform copayments or coinsurance rates have the advantage of simplicity in their administration. Moreover, these can be combined with total annual copayment ceilings (like in Norway) to protect high-risk individuals.

1.3 Copayments and income

In the presence of a financial crisis, individuals on average have less income. We may wonder whether, as a result, governments may want to rely more or less on copayments and/or coinsurance rates (e.g. a fixed proportion of expenditure). We argue below that the copayment is likely to decrease based on the theoretical analysis by Jacob and Lundin (2005).

⁴ The gradient is, in particular, observed for specialist visits and for family doctor visits (where the opposite may be observed).

They develop a political-economy model of the median voter type where individuals differ in income, there is one (uniform) coinsurance rate which is identical for everyone and health care is financed through a proportional income tax. In this set-up, they investigate whether individuals with a higher income prefer a higher or a lower copayment. There are several conditions which favour a lower copayment by the poor. First, with a decreasing relative risk aversion, poorer individuals are hit more by the copayment when they are sick. Second, if the elasticity of health care to income is less than one, then i) poorer individuals spend a higher share of their income on medical care, have a higher relative risk exposure and therefore prefer a lower coinsurance; and ii) since taxation is proportional, implicitly poor individuals are subsidised by rich individuals and they will want more health insurance, which also translates into a lower coinsurance rate. In equilibrium, the optimal coinsurance rate is determined by the voter with median income. I conjecture that in the presence of a financial crisis, the median voter will have a lower income and will therefore desire a lower coinsurance rate. The result can only be reversed if the elasticity of income is above one and/or the relative risk aversion is increasing.

From a strictly financial perspective, governments may find it tempting to raise copayments at the times of a crisis. Copayments provide a way of raising funds without raising additional taxation and/or cutting on existing expenditure. Moreover, they may curb demand, thereby further reducing expenditure (though the latter reduction may be small if demand is inelastic). The point is made more formally by Smith (2005). In contrast to the traditional insurance framework, he assumes that the budget allocated to health care is fixed but that the government can choose the optimal amount of copayments. A higher copayment allows increasing funding for additional care (i.e. increases the package covered by the public sector) but, on the other hand, may deter the poor from consuming care. From a political-economy perspective (and in line with Jacob and Lundin, 2005) individuals and patients may, however, oppose copayments if they perceive an excessive exposure to higher financial risk. Moreover, at times of a crisis, individuals are already exposed to other risks (e.g. lower income, lower employment) and the introduction of copayments clearly exposes patients even further.

1.4 Copayments and the role of doctors

The above analysis and theory of optimal health insurance in the presence of moral hazard is very much consumer-oriented or demand-driven. It implicitly assumes that the health care market is like any other market and the demand for health care is completely in the patients' control. Doctors do not play any significant role in this theory: they are passive agents that provide whatever care the patient asks for. This assumption seems implausible. The patient-doctor agency relationship is key in characterising the health sector (Arrow, 1963).

What are the implications for optimal copayments? If doctors act as perfect agents for the government or the citizens, before they become ill, and provide care only when the benefits exceed the costs, overconsumption can be completely eliminated. Having doctors acting as perfect agents is equivalent to making healthcare contractible (moral hazard is removed). We know from standard insurance theory that optimal copayments should then be set to zero (Zweifer et al., 2009).

If doctors act as imperfect agents, e.g. they provide care even when the benefits are below the costs, then copayments may still play a role, though a diminished one, if the agency relationship makes the demand for health care less elastic. The doctor-patient relationship may be influenced by the asymmetry of information which characterises them and may vary with the type of treatment. A higher asymmetry of information (for example, for patients in need of complex surgery) implies that the preferences of doctors will have a more prominent role than those of the patient, so that, as a result, demand may be more inelastic to copayments (moreover, using copayments when patients are lacking information may not necessarily deter the low benefit ones). The opposite may hold when patients may have good knowledge of the treatment (for example a well-known drug). The lower asymmetry of information should translate into a more elastic demand for copayments. One possible interpretation of why inpatient care is typically less responsive to copayments than drug expenditure is that patients have different degrees of asymmetry of information.

The patient-doctor agency relation is moreover critically influenced by how doctors are paid in addition to their intrinsic motivation or altruism. If doctors are paid by fee-for-service and insurance is open-ended (with few restrictions on supply), then doctors' and patients' interests may be

aligned. After developing an illness, patients desire *ex post* as much health care as possible (as long as the marginal benefit is not negative) and doctors may have an incentive to provide it since higher amounts of care are associated with higher revenues and profits (assuming that the price is sufficiently high) in addition to patients' benefits. Such a situation may resemble the health care market in the US in the 1970-1980's which was characterised by rapid increases in health expenditures (Cutler, 2002). In such cases, the case for copayments is a stronger one conditional on the fact that patients have some bargaining power in the decision over treatment (see Ellis and McGuire, 1990, for a formal model).

The situation may look quite different in a public health insurance system with binding budget constraints, where doctors are salaried and hospitals receive fixed budgets. In such a case, patients and doctors may have conflicting interests, the patient desiring more care, and doctors trying to stay within the budget. In such cases, the supply-side incentives may be sufficiently strong to eliminate overconsumption, moral hazard and the need for copayments (Ellis and McGuire, 1990). Moreover, since care is limited, the demand for care by patients may be more inelastic because the marginal benefit from care is higher, which once more weakens the case for copayments. On the other hand, tighter budget constraints may imply larger excess demands which may increase the scope for copayments.

Hospitals' payment based on fixed budgets has been abandoned in many countries in favour of DRG pricing which rewards hospitals for each additional patient treated. DRG payment systems have the advantage of stimulating production and productivity. Whether they stimulate excessive production depends on the generosity of the tariff and the existing capacity in the system. Generous tariffs may induce some types of overconsumption and introduce a scope for copayments, or, importantly, other forms of demand containment.

It has been well recognised in the health economics literature that doctors are motivated by and do take patients' benefits into account. However, it may be argued that this can take two different forms: one is that doctors only care about health benefits, which implies that prices or copayments are not taken into account (as in Ellis and McGuire, 1990). Therefore, if patients are passive, doctors will not respond to copayments and demand will be completely inelastic. The other possibility is that

doctors are to some extent ‘utilitarian’ and take care of the utility of the patient, not only the health benefit. In such a case, doctors will respond to copayments, even if patients are completely passive. There is some empirical evidence on this issue. One empirical study suggests that the second theory is the correct one. Iizuka (2007) examines the prescription drug market in Japan. The empirical results using anti-hypertensive drugs suggest that physicians’ prescription choices are influenced by the mark-up but are also sensitive to the patient’s out-of-pocket costs. However, doctors appear more responsive to the patient’s out-of-pocket costs than their own profits. Therefore, even if patients are passive, the demand for care will respond to copayments. Lundin (2000) finds that in Sweden, patients with large out-of-pocket expenditures (compared to patients who get most of their expenses reimbursed) are less likely to have trade-name versions of drugs prescribed. In a qualitative study, Hassell et al. (2003) find that family doctors in Italy and the UK perceived the costs for the health care system to be as important as the costs for the individual patient.

2. Waiting time and other indirect rationing

In the presence of limited copayments and a lack of explicit rationing, excess demand may still arise and patients may simply be added to a waiting list: each patient has to wait a certain time before receiving treatment. Some economists have argued that in such cases, non-price rationing in the form of waiting times takes over other forms of rationing.

2.1 A demand-supply framework

How do waiting times ration demand? There are several possible mechanisms (Iversen and Siciliani, 2011; Martin and Smith, 1999). The first mechanism is that in the presence of waiting times, some patients opt for the private sector. These may be patients who can afford it (i.e. they are richer), are more impatient, have a high private benefit, hold private supplementary insurance (which reduces the price of the private sector) or are self-employed (so that days off work translate into larger income losses).

The second mechanism is that doctors alter the clinical thresholds in response to waiting and, at the margin, refer for treatment patients with that are more severely ill when waiting times are higher. This may hold for both i) family doctors, who may refer fewer patients to hospital specialists, and ii) hospital specialists, who (conditional on having a visiting patient referred from a family doctor) may decide not to add the patient to the waiting list. This mechanism is more likely when specialists and family doctors are salaried or paid by capitation and less likely under fee-for-service arrangements.

Finally, when the waiting times are high, some patients may decide to completely give up the treatment (in any of the sectors) or opt for a drug treatment which can be taken without delay. This third mechanism can be justified if patients face transaction costs to seek health care. Longer waiting times reduce the patients' expected benefit and therefore make it less likely that the benefits overcome the transaction costs. Another possibility (a behavioural one) is that patients suffer from inertia. Waiting times generate an inconvenience and make access more complex so that some patients give up treatment.

In some extreme circumstances, patients may die while waiting. This is clearly unlikely for many elective procedures like hip replacement or cataract surgery, but could potentially be an issue for patients with, for example, cardiovascular conditions or more serious conditions. At the other end, a patient's health may improve while waiting so that a treatment is no longer required (again this seems unlikely for most elective procedures).

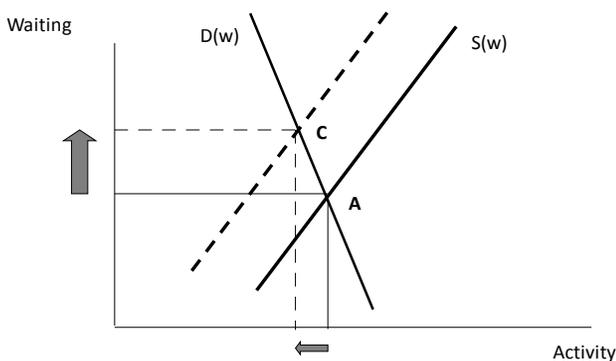
The waiting times vary extensively within publicly-funded systems. The waiting times are generally negligible (a matter of minutes or hours) for emergency care. This is because the cost of waiting is extremely high: either care is provided quickly or the patient risks a permanent impairment or dies. The waiting times are much longer (up to several months) for elective treatments, like hip replacements or cataract surgery. Although patients have positive costs of waiting, the patient can afford to wait without (hopefully) compromising their health. The waiting times vary significantly across elective procedures. Descriptive statistics across a range of countries suggest that waiting times are driven by severity or urgency criteria: for example, the waiting times for coronary bypass are much lower compared to hip replacement or cataract surgery where the

marginal disutility from waiting is arguably lower (Siciliani and Hurst, 2004).

The waiting times may not only affect the demand for care (i.e. ration demand) but also the supply of care. For example, more altruistic doctors may be willing to work harder when the waiting times are longer (Martin and Smith, 1999). If waiting times are used as targets, providers may increase the supply to avoid sanctions or penalties associated with such targets (Propper, Sutton, Whitnall and Windmeijer, 2008). Moreover, policymakers may be more willing to increase the resources when the waiting times are higher. Providers' incentives may also play a role with activity-based funding potentially associated with higher supply and lower waiting times. Activity-based financing may also stimulate the competition among providers. Whether competition translates into lower waiting times conceptually depends on the profit margin (Brekke et al., 2008). Empirical evidence from England supports the fact that competition reduces waiting times (Siciliani and Martin, 2007; Propper, Burgess and Gossage, 2008).

In summary, waiting times can be analysed within a demand-supply framework, where waiting times bring in equilibrium the demand with the supply of care (see Figure 2). Notice how an inelastic demand curve implies that an exogenous reduction in supply significantly increases the waiting times.

Figure 2. Waiting times: A demand-supply framework



Source: Own graph.

Empirically, the overall elasticity of demand for care with respect to waiting (across all treatments) has been estimated to be between -0.1 and

-0.2 in England (Martin and Smith, 1999, 2003). Therefore, a 10 per cent increase in waiting times reduces demand by 1-2 per cent. The elasticities are also reported for some specialties: for example, demand elasticity is higher (in absolute terms) for general surgery and oral surgery (respectively equal to -0.24 and -0.21) and smaller for orthopaedics (-0.07). Martin et al. (2007) report similar estimates for routine surgery and 'Ear, Nose and Throat'. They also mention that the results are broadly similar for urology and orthopaedics (though not reported in the study). Fabbri and Monfardini (2009) use survey data from Italy and find an elasticity of -0.04. Iacone et al. (2012) use a time-series approach and also find an overall demand elasticity of -0.1. The evidence on the elasticities of demand for specific treatments is very limited. Sivey (2012) provides it for cataract surgery in England and reports an elasticity equal to -0.1.

There is evidence that the supply also responds to waiting times. Early studies (Martin and Smith, 1999, 2003) which use activity both as a demand and supply measure (therefore assuming equilibrium in the market) find that the supply function is rather elastic (with an elasticity between 2 and 6, which is might be surprisingly high). Studies that have different measures of supply (patients treated) and demand (number of referrals) find that the supply elasticity is substantially lower (in the order of 0.05-0.10). It may be argued that the latter ones are more reliable since they are derived when the equilibrium assumption is relaxed.

The elasticities of demand derived in the literature can be used to predict what will happen to waiting times in the future in response to *exogenous* increases in supply. Define the variables D and S , respectively, as demand and supply, and W as waiting time. Suppose that the elasticity of demand and supply is equal to -0.2 and 0.1, respectively. We can write the model as:

$$\log(D) = \log(N) - 0.2\log(W)$$

$$\log(S) = \log(B) + 0.1\log(W)$$

where N is an exogenous demand shifter (for example related to need) and B is an exogenous supply shifter (for example related to the number of hospital beds). Assuming an equilibrium between demand and supply, i.e. $\log(D) = \log(S)$, we have that $d\log(w)/d\log(B) = -3.3$. There-

fore, a 10 per cent *exogenous* reduction in supply (B) increases waiting times by 33 per cent.

If the elasticity of demand is instead -0.1 , or the elasticity of supply is zero, then $d \log(w)/d \log(B) = -5$: a 10 per cent reduction in supply increases the waiting times by 50 per cent. Therefore, a more inelastic demand or a more inelastic supply implies that exogenous reductions in supply increase the waiting times by a larger amount.

2.2 *Optimal waiting times*

If waiting times have a rationing role to bring the demand in line with the budget allocated to the health sector, we may wonder how they should optimally be set. Gravelle and Siciliani (2008a) develop a theoretical model where a utilitarian government or public agency (which gives the same weight to different patients) has to allocate a given budget to different treatments (cataracts, bypasses, hernia, hysterectomy etc). By assumption, the public agency can only deter patients from seeking treatment in the public system by raising the waiting times. In such a context, setting the waiting times is equivalent to allocating a different supply of treatments: higher waiting times correspond to a lower supply. Moreover, by assumption, the waiting time varies across treatments but not within treatments.

The key insights of the analysis are that: i) the optimal waiting time is lower when the marginal disutility of waiting is higher; and ii) it is higher when the elasticity of demand to waiting times is higher. The first result is intuitive: if patients' costs of waiting are higher, the waiting times should be lower. The explanation for the second result is in line with the optimal copayments theory: if demand for care is more elastic to waiting, a longer wait (i.e. a higher copayment) is more effective in reducing consumption, and should therefore be set at higher levels. The two effects reinforce each other if a lower marginal disutility from waiting implies a more elastic demand or goes in opposite directions if it implies a less elastic demand. Moreover, the first effect will be reinforced if patients with a higher disutility from waiting are given a higher (equity) weight in the welfare function as compared to patients with a lower disutility from waiting.

We now discuss whether the optimal waiting time derived by Gravelle and Siciliani (2008a) can be implemented in practice. They show that under a fixed budget (F), the optimal waiting time for treatment i , defined by w_i , is equal to $w_i = \frac{Lc_i}{k_i} \varepsilon_i$ where c_i is the cost of treatment, ε_i is the (absolute value of the) elasticity of demand with respect to waiting, k_i is the marginal disutility from waiting, and L is the lagrangian multiplier associated with the budget constraint F (the fixed budget allocated to health care). This can be associated with the excess cost of raising public revenues through distortionary taxation, as discussed below in Section 3.4. As an illustrative example, we focus on cataract surgery. This is because we have empirical estimates on cataract surgery from Sivey (2012) and the elasticity of demand with respect to waiting time is equal to -0.1 . The cost of cataract surgery has been surveyed in the study by Fattore and Torbica (2008). For England, they suggest a cost of GBP 425 at 2005 prices. Propper (1990, 1995) uses a stated-preferences methodology to infer individuals' willingness to pay for reductions in waiting times. Based on a representative sample of the English population, she finds that the value of a reduction of a month on a waiting list for non-urgent treatment is GBP 35-40 (at 1987 prices). Bishai and Lang (2000) estimate that an average cataract patient would be willing to pay between USD 24 and USD 107 (at 1992 prices) for a reduction in waiting time of one month. Propper, Burgess and Gossage (2008) use a marginal disutility of waiting one month which is equal to about GBP 95 at 2003 prices (EUR 120). Suppose therefore that we set the marginal disutility of waiting one extra month equal to GBP 95. Assume further that L , i.e. the lagrangian multiplier associated with the budget constraint, is equal to 1.5. Note that this implies that the costs are inflated by 50 per cent. Using the optimal waiting time rule $w_i = \frac{Lc_i}{k_i} \varepsilon_i$ and substituting the relevant data, we have a waiting time equal to $w = (425 \times 1.5 \times 0.1)/95 = 0.67$ months. A lower marginal disutility of waiting (for example GBP 50) would instead imply a waiting time of $w = (425 \times 1.5 \times 0.1)/50 = 1.25$ months.

2.3 *Waiting time prioritisation*

Rationing by waiting can occur across different treatments (hip replacement, cataract etc), as mentioned above, but also within treatments. Waiting times vary extensively within the same procedure (for example, the waiting times for hip replacement in England in 2001 were on average 248 days with a standard deviation of 174 days, a minimum of zero and a maximum of more than three years; Laudicella et al., 2012). Doctors prioritise patients on the list within a given treatment based on urgency, severity and other criteria. Some governments have introduced explicit waiting-time prioritisation. These can involve simple methods (patients waiting less than 30, 90 and 365 days, like in Australia) or more sophisticated ones based on points-based scoring systems (like in Canada and New Zealand).⁵ In Norway, a policy has recently been implemented where patients are entitled to individualised maximum waiting time guarantees. The prioritisation criteria include severity, effectiveness of treatment, and cost in relation to the expected outcomes (Askildsen et al., 2011).

The basic idea behind prioritization policies is that waiting times should vary among patients based on observable patient characteristics. Whether such prioritisation schemes are welfare improving has been investigated in the theoretical analysis by Gravelle and Siciliani (2008b). They show that waiting-time linear prioritisation schemes based on observable characteristics are generally welfare improving. They derive this result in a set up where some dimensions of benefit are observable (for example, some dimension of health benefit) while others are not (for example a private evaluation of the benefit). Note that if benefit were perfectly observable, then it would be optimal to set the waiting times to zero, treat patients with a high benefit without delay and refuse treatment to patients with a low benefit. Gravelle and Siciliani (2008b) show that this result also holds when benefits are imperfectly observable. Even if some patients

⁵ As an example, consider the points-scheme used in New Zealand (a few years ago) for cataract patients: patients with 'lens-induced glaucoma' are assigned 71-90 points; those with 'cataract extraction required in order to treat posterior segment disease' are assigned 51-70 points. All other patients receive up to 50 points according to 'visual acuity score' (max 5 points), 'clinical modifiers' (max 5 points), 'severity of visual impairment' (max 10 points), 'ability to work, give care, live independently' (max 5 points) and 'other disability' (max 5 points). Patients with the maximum number of points wait four weeks, while patients with only 20 points wait six months. Patients scoring less than 20 are 'deferrable'.

with a high private benefit (and a low observable benefit) are not receiving treatment when it would be optimal to do so, the benefits from eliminating waiting times overcome the costs of misallocating some patients. The basic idea is that if doctors can observe some dimensions of benefit, even imperfectly, it is generally worth explicitly rationing care as opposed to by waiting times (implicitly). In summary, rationing by waiting time is generally inefficient. It imposes costs to patients which are not recovered by anyone else. Rationing on basis of observables reduces the deadweight losses from waiting. Some countries like New Zealand are moving in that direction. Patients with a low benefit are not added to the waiting list and are under “active care and review”, though if their condition deteriorates they may become eligible for treatment.

The assumption that waiting times are completely inefficient is a simplifying assumption. In practice, some waiting times can reduce the costs through a lower idle capacity and therefore a more efficient use of capacity. However, these tend to be exhausted quite rapidly. The study by Siciliani et al. (2009) suggests that at the sample mean of six-months waiting, longer waiting times do not reduce the costs, at most they increase them (maybe due to a misprioritisation of the waiting list or other costs).

2.4 Waiting times versus copayments

The previous sub-sections have investigated the role of waiting times in containing demand. The previous section (Section 1) focussed on copayments as a mechanism for containing demand. Therefore, we may wonder whether copayments and waiting times should simultaneously be used by policy makers, or whether one instrument dominates the other. Gravelle and Siciliani (2008c) show that even if a utilitarian government could choose between rationing by waiting or rationing by price, the government would always opt for prices (i.e. copayments) and set the waiting times to zero. Here, ‘zero’ waiting should be interpreted as ‘low’ levels of waiting, i.e. the minimum waiting time needed to avoid excess idle capacity (the benefits from waiting in terms of reducing idle capacity are not explicitly included in the model). This result once more arises because waiting times are a relatively inefficient way of rationing. They impose costs for patients which are not recovered by anyone else. This is in contrast to copayments which do not only contain demand but also generate

revenues which can be returned to the patient in terms of a lower premium.

Although we observe the use of copayments and coinsurance rates for hospital care, doctor's visits and drugs, we observe waiting times only for hospital care and doctor's visits but not necessarily for drugs. The patient may have to wait to see the doctor, but once the doctor has seen the patient and prescribed the drug, there is no possibility of further delaying treatment. This is in sharp contrast to elective care: once the specialist has visited and assessed the patient and suggested that surgery is needed, most of the costs associated with the treatment still need to be sustained by the provider. Therefore, there may be an incentive to postpone treatment and add the patient to the list.

At times of a crisis, it is likely that governments will cut budgets and expenditure allocated to health care. The simple simulations based on an inelastic demand suggest that such reductions will translate into significant increases in waiting times. Such increases in waiting times will help bring the market into a new equilibrium which is compatible with the reduced budget. However, it will also impose significant waiting-time costs on all inframarginal patients (those who ultimately receive the treatment). Policymakers may then consider alternative ways of rationing. One possibility is to encourage more explicit rationing. Thus, patients with a low benefit can be deterred from treatment without increasing the waiting times for all other patients (I return to this policy in the next section). There is a large empirical literature on geographical variations (Phelps, 2000). This is normally interpreted as evidence of unnecessary care. An interesting challenge would be to encourage explicit rationing towards such unnecessary care. That would help keep expenditures low without having a significant effect on patients' health.

2.5 Other forms of non-price rationing

So far, this section has focussed on waiting times as a form of non-price rationing. However, waiting times are not the only form of non-price rationing. Besley and Coate (1991) focus more broadly on quality. They argue that a government that cannot directly control expenditure may find it optimal to distort quality downwards to shift some rich patients to the private sector. This also induces a form of redistribution between rich and

poor individuals where the rich pay twice, once for the public system through the taxes (that they do not use) and then for the private sector. Gravelle and Siciliani (2008c) also investigate the optimal provision of quality in the presence of moral hazard and copayments. It suggests that a policy mix of copayments and quality distortions may be optimal. Epple and Romano (1996) also provide a political economy argument which suggests that, in the presence of a private sector, health care or its quality in the public sector will be underprovided. This arises because the median voter has an income which is below the median. The idea is that in the presence of a private sector, both the rich and the poor favour a lower quality, although for very different reasons: the rich because they already receive care in the private sector, the poor because they would like to save.

These studies tend to predict an underprovision of quality of care in the public sector. However, it is important to distinguish between two key dimensions of quality: clinical quality and amenities (quality of the room, having to share a room etc). Although the amenities are probably lower in the public sector than in the private one, it is less obvious that the same applies to clinical quality (Barros and Siciliani, 2011). Indeed, raising clinical quality is a key objective at the core of many policy developments in several publicly-funded systems. On the other hand, lower amenities are likely to affect demand by deterring some patients from seeking care in the public sector.

3. Explicit rationing

The above discussion on the role of waiting times and copayments has emphasised the drawbacks of both types of rationing. Copayments reduce the benefits from insurance. Waiting times impose costs on patients that are mostly not recovered (except for some reduction in idle capacity). From an economics point of view, a better rationing mechanism is one where patients are evaluated according to their benefits and costs, and patients with the highest benefit-cost ratio receive treatment compatibly with a given budget constraint. The waiting times and copayments can therefore be set to zero for patients receiving treatment. Patients with a low benefit-cost ratio are not provided with the treatment.

3.1 Rationing across and within treatments

A key feature of explicit rationing is the refusal of treatment to some patients. In this respect, it is useful to distinguish between rationing ‘across’ treatments and rationing ‘within’ treatments. To some extent, rationing across treatments is easier to implement. The public insurer can draw a list of treatments that are not covered. For example, certain types of cosmetic and plastic surgery or dental care can be excluded. Insurers who are more proactive in making such lists can help contain excess demand. Drawing a ‘negative’ list (of treatments excluded) is possibly simpler than drawing a ‘positive’ list (of treatments included), the latter being longer than the former. Agencies like the National Institute of Clinical Excellence (NICE) can also help suggest which treatments should be excluded or included in the package of care covered by the public insurer.

One problem with rationing ‘across’ treatments is that there is only a minority of treatments that it is not optimal to provide to any patient. The more typical case is of rationing ‘within’ a treatment where it is optimal to treat some patients but not others. Rationing explicitly ‘within’ a treatment is more complex. This can potentially be handled by the insurer by explicitly providing detailed allocation rules for each treatment based on need, severity, expected benefit, costs, equity, and other criteria. Although outlining these principles is simpler in general, it is rather costly to develop precise guidelines for each individual treatment. This is why, in practice, the decision is delegated to the doctor, who will have the double role of providing the care to the patient and also of making the difficult decisions of who should receive the treatment and who should not.

3.2 The role of doctors

If doctors can explicitly ration patients in a costless way, then (as already argued above) welfare will be higher as compared to other forms of (price or waiting-time) rationing. Although doctors do to some extent explicitly ration care, the long waiting lists are evidence that explicit rationing is limited in some public systems. The reason may be that explicitly rationing patients is in itself costly for doctors. It is the doctor who will have to tell the patient that, despite some positive (but small) benefits, the patient cannot be treated; alternatively, it may be that the patient is not sufficiently severely ill (as for some elective treatments); or, that the patient is so

severely ill that treatment would provide very little benefit. The tighter is the rationing, the higher will be the number of patients who may argue against it, insisting that they need treatment. When care is refused, the doctor may be held responsible and liable by the patient for taking the wrong decision, i.e. turning the patient down when she should have been treated. Patients may also decide to make a formal complaint to the hospital, the health authority or other institutions. Moreover, stricter rationing implies that doctors need to spend more time assessing patients' benefit in order to exclude those with a lower benefit (Siciliani, 2007). From the doctor's point of view, it may be less costly to add the patient to the waiting list. If explicit rationing is costly for doctors, there may be scope for combining explicit rationing with waiting time rationing. At one extreme, explicit rationing will generate excessive costs for the doctors; at the other extreme, rationing exclusively by waiting will generate excessive costs for the patients. It may then be optimal to combine explicit rationing by the doctor with implicit rationing by waiting (Siciliani, 2007).

There are two types of doctors that play a role in explicit rationing: the family doctor and the hospital specialist. Some countries have a gatekeeping system where family doctors potentially have a more prominent role: patients need to visit and obtain a referral from the family doctor to access specialist or hospital care. Other countries do not have a gatekeeping system and patients can directly access specialist care. It is typically countries with lower spending that are concerned about excessive numbers of referrals and therefore have a gatekeeping system. However, if family doctors simply refer all patients to specialists, their rationing role becomes void. It is only if they refuse treatment to some patients that the number of referrals will be contained. The degree of rationing exerted by family doctors will also depend on the financial incentives and whether referrals have financial implications for them. In England under the fundholding scheme during 1991-1999, family doctors could choose to have a budget and pay for the costs of certain types of elective hospital surgery. They could retain any surplus. When fundholding was abolished, the activity increased by 3.5 to 5 per cent (Dusheiko et al., 2006). The gatekeeping role of family doctors is therefore reinforced when they are made financially responsible for hospital care. Recent reforms known as 'Practice based commissioning' have a similar rationale. The idea of gatekeeping doctors has also been adopted in the US within Health Maintenance

Organisations, which differs from more traditional insurance where patients can seek care directly from a specialist.

3.3 Estimates of the marginal benefit

At times of a crisis, it is likely that reductions in GDP will lead to lower expenditures for the health sector. The seminal paper of Newhouse (1977) used a cross-section of data among OECD countries and suggested GDP to be the key predictor of health expenditure. Following studies have used a panel data or a time-series approach and have generally confirmed such findings, although causality is more difficult to establish as the time dimension increases (more technically, the variables may be cointegrated; see for example Gerdtham and Löthgren, 2000).

If health expenditure is reduced, policy makers need to ration where the marginal benefit from expenditure is lowest and maintain expenditure where the marginal benefit is highest. This would require proper estimates of the marginal benefit for each treatment (per unit of cost). The development of such estimates seems unlikely in the short run. One interesting development which goes in this direction is the estimation of marginal benefits by programme budgeting area in England (Martin et al., 2008). Expenditure has been split into 23 broad programmes of care (for example cancer, circulation problems, respiratory problems, etc) and data on both outcomes and expenditure is available for about 300 Primary Care Trusts. The variation in expenditure and outcomes allows us to estimate the marginal benefit from extra spending. Endogeneity is clearly an issue, e.g. more spending allocated where health outcomes are worse, but this can be dealt with in an instrumental-variable approach. The results suggest in 2004/2005 for cancer a 1 per cent increase in expenditure per head (GBP 0.751) which is associated with a 0.378 per cent reduction in life years lost (0.021 days) and implies that one life year would cost GBP 13 137. For circulation problems, a 1 per cent increase in the expenditure per head (GBP 1.22) is associated with a 1.4 per cent reduction in life years lost (0.056 days) and implies that one life year would cost GBP 7 979. Martin et al. (2009) update the results for 2006/2007 and find that the marginal cost of a life year saved is GBP 15 387 for cancer, GBP 9 974 for circulation problems, GBP 5 425 for respiratory problems, GBP 21 538 for gastro-intestinal problems and GBP 26 429 for diabetes. The

figures can be adjusted to take into account Quality-Adjusted Life-Years (QALY). Martin et al. (2008) suggest that the cost of a QALY is GBP 19 070 for cancer and GBP 11 960 for circulatory problems.

3.4 Estimating the tightness of the budget constraint

As argued above, rationing is costly for *doctors*. Such costs are higher the larger is the difference between demand and supply, which the allocated budget can afford. This paragraph argues that the amount of rationing in publicly-funded systems is significant. In the absence of distortionary effects from raising taxes, a utilitarian government would set the optimal level of spending such that the marginal benefit from treatment b is equal to its marginal cost c so that, at the optimum, we have $b = c$. In the presence of distortionary taxation, an extra Euro raised from taxation generates distortions equal to L (for example in the labour market). In public economics, this is normally known as the opportunity cost of public funds. Costs therefore need to be multiplied by $(1 + L)$ and the optimal allocation rule is such that $b = (1 + L)c$: the marginal cost is now inflated. A different interpretation for L is that it represents the lagrangian multiplier associated with the budget constraint. Suppose that as part of the political process, a government allocates a given budget to the health sector. Then, L can be interpreted as the multiplier which is compatible with that budget. Tighter budget constraints are associated with larger values of L . Consider the following illustrative example. The study by Martin et al. (2008), discussed in the previous paragraph (which exploits variations in expenditure by Primary Care Trusts), suggests that in England, the marginal cost of one QALY, c , is equal to GDP 19 000 for cancer. Suppose that the value of a QALY, b , is equal to GDP 30 000 (Donaldson et al., 2011). By comparing the two figures, we obtain an estimate of $(1 + L) = b/c = 30000/19000 = 1.67$. This suggests that the costs need to be inflated by 67 per cent to make expenditure compatible with the budget constraint. For circulatory problems, Martin et al. (2008) estimate an even lower marginal cost of one QALY equal to GDP 12 000, which by the same computations implies that $(1 + L) = 2.5$. In this case, the costs are inflated by 150 per cent. A lower benefit of a QALY would imply a lower L . Suppose that we assume a benefit from a QALY equal to GDP 20 000. Then, the estimation of $(1 + L)$ is now, respectively,

equal to $20000/19000=1.05$ and $20000/12000=1.67$ for cancer and circulation problems. A value of a QALY of 40 000 would instead imply an estimate of $40000/19000=2.10$ and $40000/12000=3.3$ for cancer and circulation problems. These simple examples emphasise that the amount of rationing in publicly-funded health systems is potentially significant: positive and higher values of L imply that the budget constraint is more binding and care has to be rationed up to the point where the marginal benefit is strictly above the marginal cost.

OECD countries vary substantially in public health spending, both when measured in per-capita terms or as a percentage of GDP. Some of these countries do not differ dramatically in terms of demographics (on the demand side) but can vary substantially in health spending. For example, public spending in 2009 per capita was 1 608 in Spain, 2 178 in the UK and 2 921 in Norway in dollars at USD 2 000 (Purchasing Power Parity – GDP deflator). Differences in spending reflect differences in supply that will translate into differences in excess demand (although note that differences in expenditures may also be due to differences in the labour force prices of health care employees, in addition to supply).

3.5 Combining explicit rationing with waiting time and price rationing

Finally, explicit rationing can be used in combination with waiting time and price rationing. For a given excess demand, waiting time and price rationing may reduce demand and require less explicit rationing. To illustrate this point, simply suppose that patients differ uniformly in benefit b so that patients with the lowest benefit have zero benefit, and patients with the highest benefit have a benefit equal to B . Suppose that patients are explicitly rationed (with no waiting time and copayments): patients with a benefit above x receive treatment and others do not. In the absence of price and waiting time rationing, x patients will need to be rationed. Now suppose that the government introduces a small copayment p and a small wait w (and, moreover, the marginal disutility of waiting is normalised to one). Then, patients with benefit $b < w + p$ will not demand treatment: only $x - (w + p)$ patients are explicitly rationed. In summary, patients whose benefit falls within $(0, w + p)$ are rationed by waiting and by copayments; patients whose benefit falls within $(w + p, x)$ are explicitly rationed. An increase in copayments or waiting times reduces the

amount of explicit rationing necessary to bring demand in equilibrium with supply. As already argued above, waiting times impose costs on patients, copayments reduce the benefits from insurance, and explicit rationing is costly for doctors. A combination of the three rationing mechanisms is possible. Whether it is desirable depends on the relative benefits and costs of each type of rationing.

4. Discussion and conclusions

We have discussed the relative merits of three different forms of rationing. We have started by price rationing. Optimal insurance theory (under moral hazard) suggests that optimal copayments and coinsurance rates should be set to balance protection from the risk of medical expenses against the tendency towards overconsumption when patients need medical care. It also suggests that the optimal copayments should be higher when the demand for health care is more elastic because the deadweight loss from overconsumption is higher. The prediction seems consistent with what we observe in several countries: copayments are generally higher for drugs, where demand is arguably more elastic, and lower for inpatient care where demand is less elastic.

Given the current economic climate, a critical question is whether copayments should increase or decrease at times of lower growth and GDP. The standard insurance framework suggests that a utilitarian government (or a government driven by the preferences of the median voter) would reduce copayments. This arises because individuals value more coverage from the risk of healthcare payments when income is lower, despite the larger premium which arises from the lower copayment. Within publicly-funded systems, larger premia will imply more taxes to be raised. However, if there are strict limits for the extent to which taxes can be raised, governments may instead be quite tempted to increase copayments to both finance health expenditure without raising taxation and to curb consumption. Equity and risk exposure will suffer as a result, however.

Copayments play a limited role in many publicly-funded systems, especially for inpatient care. This implies that excess demand is often dealt with by adding patients to the list. Normative theory suggests that the waiting times should be longer when the elasticity of demand is higher

and lower when the marginal disutility from waiting is higher. The first result is in line with optimal copayment theory. A more elastic demand implies that a marginal increase in waiting time is more effective in containing demand. The result that a higher marginal disutility should lead to lower waiting times is consistent with doctors' prioritisation: patients who suffer most from waiting should wait less.

The empirical studies suggest that demand is inelastic. Simple simulations show that, assuming a demand elasticity of -0.2 and a supply elasticity of 0.1, a 10 per cent reduction in health expenditure implies an increase in waiting times by 33 per cent. This suggests that other policies may need to be put in place if such dramatic increases are to be avoided. More explicit rationing may help contain increases in waiting times.

Explicit rationing has the potential of maximising patients' and overall welfare as compared to price and waiting time rationing. Copayments reduce the benefits from insurance, and waiting times impose time costs on patients. Both limitations can be overcome by an explicit rationing system where patients are evaluated according to their benefits and costs and patients with the highest benefit-cost ratio receive treatment that is compatible with a given budget constraint. Under explicit rationing, patients have full insurance and no disutility from waiting.

A key feature of explicit rationing is the refusal of treatment to some patients. Given its potential benefits compared to rationing by waiting or by copayments, it is to some extent surprising that it is not more extensively used. The reason may be that explicit rationing can be costly for the doctors which may limit its use for elective care. Under explicit rationing, doctors need to spend more time assessing patients' benefit in order to exclude those with a lower benefit; they may be held directly responsible by the patient for taking the wrong decision; patients may argue against the doctor's decision, insisting that they need treatment and make a formal complaint. Doctors' costs from rationing will be higher the larger is the gap between demand and supply. In many cases, the doctors' explicit rationing role is made more difficult by the lack of precise guidance and a different interpretation can be given to general criteria such as need, severity, benefits and costs. The development of more precise guidelines is an interesting area for future policy development. Some countries (like Canada, New Zealand, Norway and Sweden) have been proactive in developing such guidelines. Policies that make family doc-

tors (such as practice-based commissioning) more financially responsible for hospital referrals may also help share the costs from explicit rationing between family doctors and specialists.

References

- Arrow, K. (1963), Uncertainty and the welfare economics of medical care, *American Economic Review* 53, 941-973.
- Askildsen, J.E., Holmås, T.H. and Kaarbøe, O.M. (2011), Monitoring prioritization in the public health care sector by use of medical guidelines: The case of Norway, *Health Economics* 20, 958-970.
- Barros, P.P. and Siciliani, L. (2011), Public-private interface, in M. Pauly, T. McGuire and P.P. Barros (eds), *Handbook in Health Economics* 2, Elsevier, Amsterdam.
- Besley, T. and Coate, S. (1991), Public provision of private goods and the redistribution of income, *American Economic Review* 81, 979-984.
- Bishai, D.M. and Lang, H.C. (2000), The willingness to pay for wait reduction: The disutility of queues for cataract surgery in Canada, Denmark, and Spain, *Journal of Health Economics* 19, 219-230.
- Brekke, K., Siciliani, L. and Straume, O.R. (2008), Competition and waiting times in health care markets, *Journal of Public Economics* 92, 1607-1628.
- Cremer, H. and Pestieau, P. (1996), Redistribution taxation and social insurance, *International Tax and Public Finance* 3, 281-298.
- Cutler, D. (2002), Equality, efficiency, and market fundamentals: The dynamics of international medical care reform, *Journal of Economic Literature* 40, 881-906.
- Donaldson, C. et al. (2011), The social value of a QALY: Raising the bar or barring the raise?, *BMC Health Service Research* 11, 8.
- Dusheiko, M., Gravelle, H., Jacobs R. and Smith, P.C. (2006), The effect of financial incentives on gatekeeping doctors: Evidence from a natural experiment, *Journal of Health Economics* 25, 449-478.
- Ellis, R.P. and McGuire, T.G. (1990), Optimal payment systems for health services, *Journal of Health Economics* 9, 375-396.
- Epple, D. and Romano, R. (1996), Public provision of private goods, *Journal of Political Economy*, 104, 57-84.
- Fabbri, D. and Monfardini, C. (2009), Rationing the public provision of healthcare in the presence of private supplements: Evidence from the Italian NHS, *Journal of Health Economics* 28, 290-304.
- Fattore, G. and Torbica, A. (2008), Cost and reimbursement of cataract surgery in Europe: A cross-country comparison, *Health Economics* 17, S71-S82.
- Feldman, R. and Dowd, B. (1991), A new estimate of the welfare loss of excess health insurance source, *American Economic Review* 81, 297-301.
- Gerdtham, U.G. and Löthgren, M. (2000), On stationary and cointegration of international health expenditure and GDP, *Journal of Health Economics* 19, 461-475.
- Goldman, D. and Philipson, T.J. (2007), Integrated insurance design in the presence of multiple medical technologies, *American Economic Review* 97, 427-432.

- Gravelle, H. and Siciliani, L. (2008a), Ramsey waits: Allocating public health service resources when there is rationing by waiting, *Journal of Health Economics* 27, 1143-1154.
- Gravelle, H. and Siciliani, L. (2008b), Is waiting-time prioritisation welfare improving?, *Health Economics* 17, 167-184.
- Gravelle, H. and Siciliani, L. (2008c), Optimal quality, waits and charges in health insurance, *Journal of Health Economics* 27, 663-674.
- Hassell, K., Atella, V., Schafheutle, E.I., Weiss, M.C. and Noyce, P.R. (2003), Cost to the patient or cost to the healthcare system? Which one matters the most for GP prescribing decisions?, *European Journal of Public Health* 13, 18-23.
- Iacone, F., Martin, S., Siciliani, L. and Smith, P.C. (2012), Estimating a dynamic model of waiting times with time series data, *Applied Economics* 44, 2955-2968.
- Iizuka, T. (2007), Experts' agency problems: Evidence from the prescription drug market in Japan, *RAND Journal of Economics* 38, 844-862.
- Iversen, T. and Siciliani, L. (2011), Non-price rationing and waiting times, in S. Glied and P. Smith (eds.), *Oxford Handbook of Health Economics*, Oxford University Press, Oxford.
- Jacob, J. and Lundin, D. (2005), A median voter model of health insurance with ex post moral hazard, *Journal of Health Economics* 24, 407-426.
- Laudicella, M., Siciliani, L. and Cookson, R. (2012), Waiting times and socioeconomic status: Evidence from England, *Social Science and Medicine* 74, 1331-1341.
- Lundin, D. (2000), Moral hazard in physician prescription behavior, *Journal of Health Economics* 19, 639-662.
- Manning, W.G. and Marquis, M.S. (1996), Health insurance: The tradeoff between risk pooling and moral hazard, *Journal of Health Economics* 15, 609-639.
- Manning, W.G., Newhouse, J., Duan, N., Keeler, E.B., Leibowitz, A. and Marquis, M.S. (1987), Health insurance and the demand for medical care: Evidence from a randomized experiment, *American Economic Review* 77, 251-277.
- Martin, S., Rice, N., Jacobs, R. and Smith, P.C. (2007), The market for elective surgery: Joint estimation of supply and demand, *Journal of Health Economics* 26, 263-285.
- Martin, S., Rice, N. and Smith, P.C. (2008), Does health care spending improve health outcomes? Evidence from English programme budgeting data, *Journal of Health Economics* 27, 826-842.
- Martin, S., Rice, N. and Smith, P.C. (2009), The link between healthcare spending and health outcomes for the new English primary care trusts, Report for QQUIP, The Health Foundation, London.
- Martin, S. and Smith, P.C. (1999), Rationing by waiting lists: An empirical investigation, *Journal of Public Economics* 71, 141-164.
- Martin, S. and Smith, P.C. (2003), Using panel methods to model waiting times for National Health Service surgery, *Journal of the Royal Statistical Society*, 166/Part 2:1-19.
- McGuire, T. (2011), Demand for health insurance, in M. Pauly, T. McGuire and P.P. Barros (eds.), *Elsevier Handbook in Health Economics* 2, Elsevier, Amsterdam.
- Newhouse, J.P. (1977), Medicare expenditure: A cross-national survey, *Journal of Human Resources* 12, 115-125.
- Nyman, J.A. (1999a), The economics of moral hazard revisited, *Journal of Health Economics* 18, 811-824.

- Nyman, J.A. (1999b), The value of health insurance: The access motive, *Journal of Health Economics* 18, 141-152.
- Nyman, J.A. (2012), The value of health insurance, in A. Jones (ed.), *Elgar Companion to Health Economics*, Edward Elgar Publishing, Cheltenham.
- Pauly, M.V. (1968), The economics of moral hazard: Comment, *American Economic Review* 58, 531-537.
- Phelps, C.E. (2000), Information diffusion and best practice adoption, in A.J. Culyer and J.P. Newhouse (eds.), *Handbook of Health Economics*, Elsevier, Amsterdam.
- Propper, C. (1990), Contingent valuation of time spent on NHS waiting lists, *Economic Journal* 100, 193-200 (Conference Supplement April 1990).
- Propper, C. (1995), The disutility of time spent on the United Kingdom's National Health Service waiting lists, *Journal of Human Resources* 30, 677-700.
- Propper, C., Burgess, S. and Gossage, D. (2008), Competition and quality: Evidence from the NHS internal market 1991-9, *Economic Journal* 118, 138-170.
- Propper, C., Sutton, M., Whitnall, C. and Windmeijer, F. (2008), Did 'targets and terror' reduce waiting times in England for hospital care?, *B.E. Journal of Economic Analysis & Policy* 8(2), (Contribution), Article 5.
- Siciliani, L. (2007), Optimal contracts for health services in the presence of waiting times and asymmetric information, *B.E. Journal of Economic Analysis & Policy* 40, 1-25.
- Siciliani, L. (2012), Rationing of demand, in T. Culyer (ed.), forthcoming in *Encyclopedia of Health Economics*, Section 8.
- Siciliani, L. and Hurst, J. (2004), Explaining waiting times variations for elective surgery across OECD countries, *OECD Economic Studies* 38, 96-122.
- Siciliani, L. and Iversen, T. (2012), Waiting lists and waiting times, in A. Jones (ed.), *Elgar Companion in Health Economics*, Edward Elgar, Cheltenham.
- Siciliani, L. and Martin, S. (2007), An empirical analysis of the impact of choice on waiting times, *Health Economics* 16, 763-779.
- Siciliani, L., Stanciole, A. and Jacobs, R. (2009), Do waiting times reduce hospitals' costs?, *Journal of Health Economics* 28, 771-780.
- Sivey, P. (2012), The effect of waiting time and distance on hospital choice for English cataract patients, *Health Economics* 21, 444-456.
- Smith, P.C. (2005), User charges and priority setting in health care: Balancing equity and efficiency, *Journal of Health Economics* 24, 1018-1029.
- Wagstaff, A. and van Doorslaer, E. (2000), Equity in health care finance and delivery, in A.J. Culyer and J.P. Newhouse (eds.), *Handbook of Health Economics* 1, Elsevier, Amsterdam.
- Zeckhauser, R. (1970), Medical insurance: A case study of the trade off between risk spreading and appropriate incentives, *Journal of Economic Theory* 2, 10-26.
- Zweifel, P., Breyer, F. and Kifmann, M. (2009), *Health Economics*, Springer, Berlin.

Comment on Siciliani: An economic assessment of price rationing versus non-price rationing of health care

Mickael Bech*

Health care expenditures will continue to rise. Luigi Siciliani's paper provides a stringent and very readable overview of the contributions of economic theory and empirical analyses to the issue of the rationing of health care services. This paper provides inspiration for policymakers who face the difficult task of balancing the obvious needs for rationing and the political feasibility of the different rationing mechanisms. It will serve as a guide to policy makers, as it contains arguments with which traditions and current choice of rationing mechanisms can be benchmarked.

The starting point of the paper is the pros and cons of copayment. The paper provides examples of the current implementation of copayment which, in accordance with standard insurance theory, involves a higher copayment for services with higher own-price elasticity. Yet, there are also very good examples of how the copayments across areas are mostly determined by historical paths (Pedersen et al., 2005). These historical paths are not easily overcome, but policymakers may find help in Siciliani's paper.

As briefly emphasised in the paper, the optimal design of copayment does not only depend on own-price elasticity of one category of health care services, but also on how the demand for other types of care may be

* COHERE – Centre of Health Economics Research, Department of Business and Economics, University of Southern Denmark, mbe@sam.sdu.dk.

affected – and whether the other types of care are substitutes or complements. Here, a full health system perspective is needed to assess the pros and cons of the allocation of copayment across areas. One of the examples, discussed in some of the Nordic countries, is copayments for GPs and emergency ward visits which could be partial substitutes. GP visits and outpatient visits seem to be partial substitutes (Bech and Lauridsen, 2009), which creates a need to coordinate copayment across areas to avoid unintended and adverse shifts toward more costly health care resources. Policymakers will definitely be familiar with this discussion, but information such as cross-price elasticities is not necessarily available and there is still a great need for more empirical studies in this area to guide policy makers.

Copayment will most likely harm equity unless counterbalancing design parameters are carefully included in the copayment scheme. Caps on overall copayment may be one approach. When the copayment cap is fixed at a small amount, many patients will pass this limit, after which care is provided for free. This involves less harm to equity, but also less reduction of overconsumption. Another approach is to introduce income-tested copayment, which limits the harmful effect on equity but also involves a lower reduction of overconsumption. Furthermore, income-testing also involves that the copayment policy will become a matter of the overall tax policy in a country because it effectively affects the marginal tax rate.

The second mechanism is to have waiting time for treatment, which imposes costs on patients due to the cost of waiting. Some of these costs may be counterbalanced by the reduced costs through a more efficient use of capacity, but this argument is only valid within certain limits of waiting time, as pointed out in the paper. If the various mechanisms through which waiting lists will ration health care consumption work effectively and do not simply re-allocate health care expenditure across areas, the tax payers retain some of the patients' cost of waiting time.

Beyond the rationing effects of waiting lists, there may in some cases be clinical reasons for maintaining some waiting time before treatment. This may be especially relevant for some of the areas of orthopaedic surgery where the evidence of surgery is ambiguous and where other treatment modalities and patient efforts may be beneficial.

The last mechanism is explicit rationing, which will in theory ration treatment to the point where marginal benefit equals marginal cost. The implementation of explicit rationing may impose costs on the individual doctors who are to implement this in practice. Some of these costs may be mitigated by reformulating the clinical guidelines to support doctors in the interaction with the individual patient. However, these general guidelines do not come without a cost, and the Scandinavian countries are reluctant to follow the English example of NICE because of the cost of producing the information basis for explicit rationing.

The political cost also seems to be an influential barrier for explicit rationing. From a simple observation, it seems to be hard for politicians to explicitly turn down a treatment. Politicians have a hard time and may not dare to explain the principles of opportunity cost to the population. Even if politicians agree on the basic economic principles for rationing, we cannot prevent opportunistic politicians from breaking a highly unstable equilibrium of political agreement by declining access to a specific treatment. The explicit rationing may involve too many political costs and this may be an argument for trying to place these decisions in more bureaucratic institutions.

A final consideration could be how the three rationing mechanisms affect the public support for the public health care system. If the public support is severely harmed, the very basis for a public system may be undermined in the long run. This will probably be an argument for a mixed strategy where none of the mechanisms are used in their extremes and marginal harms to public support are balanced.

The paper concludes by summarising the relative merits of the three mechanisms. Siciliani's overview provides support for extending the use of explicit rationing relative to the other two mechanisms with convincing economics arguments. Hopefully, this paper will be read by many policymakers since it definitely contains an important contribution to the health policy debate in the Nordic countries.

References

- Bech, M. and Lauridsen, J.T. (2009), Exploring the small area variation and spatial patterns in outpatient treatments, *Health Services and Outcomes Research Methodology* 9, 177-196.

Pedersen, K.M., Christiansen, T. and Bech, M. (2005), The Danish health-care system: Evolution – not revolution – in a decentralized system, *Health Economics* 14, 41-57.

Should pharmaceutical costs be curbed?*

Kurt R. Brekke^{**}, Dag Morten Dalen^{***} and Steinar Strøm^{****}

Summary

Pharmaceuticals account for almost a fifth of total health spending in OECD-countries. Both pharmaceutical innovations and the aging of the population explain the increasing importance of pharmaceuticals in health care. Due to the importance of patent protection and insurance coverage, pharmaceutical markets are subjected to economic regulation – both on the supply-side and the demand-side. In this paper, we briefly review the Nordic pharmaceutical market, before explaining the main regulatory policy measures taken by governments in these countries. Empirical research has been undertaken to investigate regulation and competition, and we provide a review of some of the findings.

Keywords: pharmaceutical markets, pharmaceutical costs, reference pricing, price cap, health insurance.

JEL classification numbers: I11, I13, I18, L51.

* We are grateful for comments from participants at the Nordic Economic Policy Review Conference on Economics of Health Care (May 2012, Reykjavik) and at the Norwegian Health Economic Conference (May 2012, Oslo). Financial support from the Norwegian Research Council is acknowledged.

** Norwegian School of Economics, kurt.brekke@nhh.no.

*** BI Norwegian Business School, Dag.m.dalen@bi.no.

**** University of Turin, Steinar.strom@econ.uio.no.

Pharmaceuticals have become an important part of health care, both in terms of treatment outcomes and in terms of public spending. Pharmaceuticals now account for almost a fifth of total health spending in OECD-countries.¹ Both pharmaceutical innovations and the aging of the population explain the increasing importance of pharmaceuticals in health care. Cancer, high blood pressure and cholesterol, and depression are examples of diseases where pharmaceutical innovations have improved the treatment, but also triggered increased costs for public health insurance schemes in Nordic countries. In Norway, the consumption of anti-cholesterols per inhabitant increased by close to 340 percent from 2000 to 2009. Both Finland and Denmark have seen similar growth rates. The growth rates in Sweden and Iceland have been lower, but also these countries have seen a sharp increase in the consumption of anti-cholesterols.

The life-cycle of a new drug entering the market can be divided into two phases. The first phase is the one in which a patent protects the innovating company from direct competition from other companies. The patent holder has exclusive rights to produce and sell the drug. The second phase begins when the patent expires and other firms are free to produce and market the exact same – generic – drug.

The abilities of governments to control – or curb – costs in these two phases are very different. When a new drug is approved and enters the market, the main mechanism for controlling costs is by setting requirements for prescription and reimbursement (restricting the use) and by setting price-caps. In addition, parallel import of pharmaceuticals within EU restricts the ability of the innovating company to increase prices in one single country – competition between direct import and parallel import to some extent hinders third-degree price discrimination in the European market. In Sweden and Denmark, parallel import is actively used when determining the patients' copayment (see Section 3.3).

For a patented drug, cost control is closely linked to the quality of the drug relative to other treatment options for the same disease. Reducing spending on innovative drugs (phase one) involves a tradeoff with quality of health care. If a new drug enables a considerably improved treatment compared to other available drugs, curbing the costs for this patient group may be welfare-reducing. If instead the new drug is less innovative (“me-

¹ Health at a Glance (2011).

too-drug”), the cost can be curbed with therapeutic competition without hurting the patient.

With the entry of generic drugs (phase two), large cost savings can be realized if the insurance schemes are able to trigger price competition among the producers. A successful implementation of generic competition can generate a cost saving without lowering the quality of treatment.

The rest of this paper is organized as follows. In Section 1, we briefly review the Nordic pharmaceutical market. In Section 2, we continue by explaining the main regulatory policy measures used by governments in these countries. Section 3 provides a review of some of the findings in the empirical research that has been undertaken to investigate regulation and competition in these markets. We conclude the paper in Section 4 by returning to the question raised by the title of our paper.

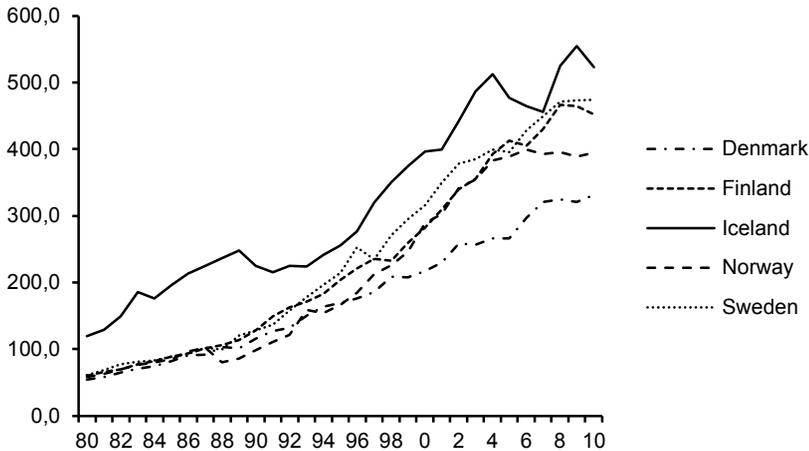
1. Nordic pharmaceutical markets

In this section, we take a closer look at the pharmaceutical market in the Nordic countries. We first describe the level and development of pharmaceutical expenditure in the national markets. Then, we consider the price and the consumption levels of pharmaceuticals in order to explore sources of variation in pharmaceutical spending across countries.

1.1 Pharmaceutical sales/expenditures

The figure below shows the development of pharmaceutical expenditure per capita, measured in USD (OECD-purchasing power parity), from 1980 to 2010. There is a significant variation across the Nordic countries. Iceland has the highest level of pharmaceutical expenditure with USD 523 per capita in 2010, whereas Denmark has the lowest expenditure level with USD 331 per capita. Thus, pharmaceutical expenditure in Iceland is almost 60 percent higher than in Denmark, when making the OECD-purchasing power adjustment. We also see that Finland, Norway and Sweden experienced higher growth rates than Denmark during the 1990’s.

Figure 1. Pharmaceutical expenditure (USD-PPP) per capita, 1980-2010



Source: OECD Health Data.

The average annual growth in pharmaceutical expenditure (measured in USD-PPP) from 1990 to 2004 was as high as 12 percent for Norway and 10 percent for Sweden. The annual growth rate in Denmark was 7 percent. In Norway, pharmaceutical expenditure has been stable, or even slightly declining, since 2004.

International comparisons of consumption levels are difficult and controversial (see Almås, 2012). A simple exchange rate conversion changes the picture dramatically by turning Denmark into a country with high pharmaceutical consumption as compared to other Nordic countries. The expenditure levels in 2011, measured with euro per capita, are as follows:

Table 1. Pharmaceutical expenditure per capita. Euro. 2011

	Denmark	Finland	Norway	Sweden
Euro per capita, 2011	400	360	310	360

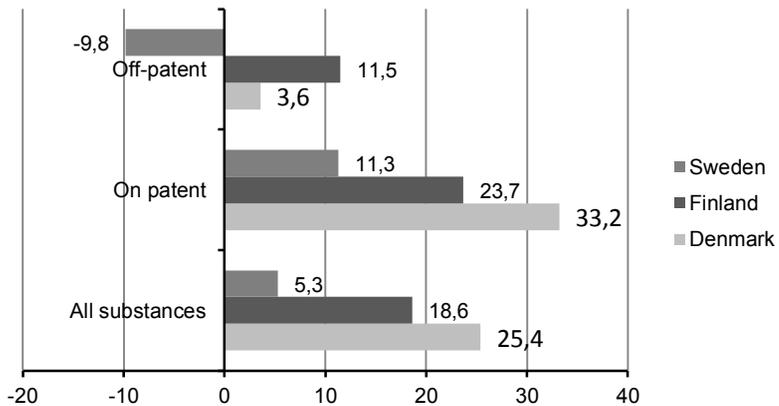
Source: LMI. Facts and figures.

The difference between the two measures in Figure 1 and Table 1 is due to the PPP-adjustment.

1.2 Prices of pharmaceuticals

Pharmaceutical expenditure (sale) is the product of prices and volumes. In this section, we consider the prices in the Nordic countries, whereas in the next section we consider the volumes. Comparing the prices of pharmaceuticals across countries is a challenge, since these products are by nature heterogeneous.² There have been a couple of recent studies on the pharmaceutical price levels including the Nordic countries. Brekke et al. (2011b) compare the prices of pharmaceuticals in Norway with nine European countries. They use a sample of the 300 most selling substances as the basis for comparison, and compute a wide set of price indices in order to measure the price levels for all substances and for various submarkets such as the on-patent and off-patent market segments. The figure below reports the price indices for the Nordic countries with Norway as the base country with a price index normalized to zero.

Figure 2. Bilateral price indices based on average substance prices at pharmacy levels, 2010



Source: Brekke et al. (2011b).

We see that the Norwegian price level tends to be the lowest among the Nordic countries. If we look at all substances in the sample, importing the Swedish price level would result in a 5.3 percent increase in the pharmaceutical expenditures in Norway assuming that the consumption is

² Danzon (1999) gives a detailed discussion of challenges related to cross-country price comparisons.

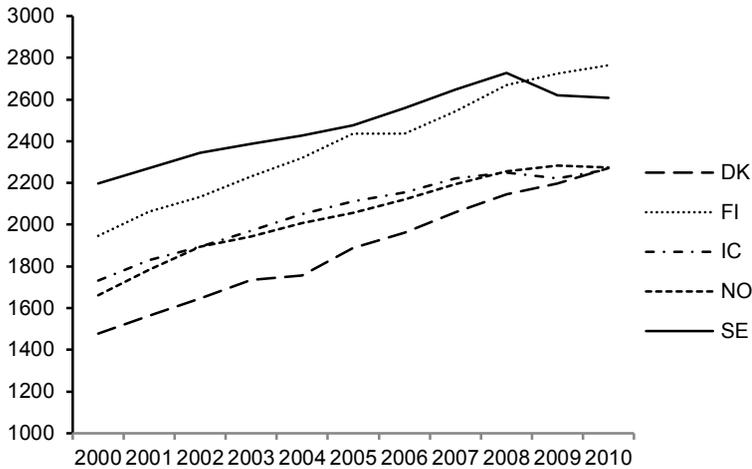
unchanged. Importing the Finnish and Danish price levels results in even higher expenditure increases of 18.6 and 25.4 percent, respectively. The price differences are higher in the on-patent market segment, while for substances in the off-patent market segment with generic competition Sweden does, in fact, have a lower price level than Norway.

Note that if the price indices show that the Norwegian consumption of pharmaceuticals would be 10 percent more expensive if using, say, Swedish prices, the reverse is not necessarily true. The reason is that we then need to replace the Norwegian consumption weights with the Swedish consumption weights, implying that although the prices are the same, the price indices would be different. In a recent report, Brekke and Holmås (2012) have computed the Swedish price indices for the on-patent market segment and contrasted these with the price indices obtained in Brekke et al. (2011b). The results show that there is a weak tendency for the base country to become cheaper, but the results are not qualitatively altered. This is also confirmed in a recent study by the Swedish regulatory body TLV written by Arnberg et al. (2012). Thus, cross-country price differences might partly explain the differences in pharmaceutical sales (expenditures) per capita between the Nordic countries.

1.3 Consumption of pharmaceuticals

The second source of differences in pharmaceutical sales across the Nordic countries is the consumption of pharmaceuticals. Since drugs are sold in different pack sizes with different strengths and formulations, we need a measure to make consumption comparable across substances. The most common measure is defined daily doses (DDDs).³ The figure below shows the consumption of pharmaceuticals (in DDDs) normalized by per 1 000 capita for the Nordic countries.

³ This is a measure developed by the World Health Organization (WHO) that allows for a comparison of consumption across products with different substances, dosages, and formulations. A DDD is based on the recommend treatment for the main indication of the specific drug. For instance, if the DDD is 20mg of a given substance, then a pack with 10 tablets of 10mg would yield a volume of 5 DDDs.

Figure 3. Consumption, DDD per 1 000 capita

Source: OECD Health Data: Pharmaceutical market.

We see that the consumption of pharmaceuticals in Denmark, Iceland and Norway is low as compared to in Finland and Sweden. Finland has the highest consumption level in 2010 with close to 2 800 DDDs per 1 000 capita. This is about 20 percent higher than the consumption level in Norway, Denmark and Iceland. The variation in the consumption of pharmaceuticals across the Nordic countries is substantial, but they have all experienced a high consumption growth with an accompanying increase in pharmaceutical costs.

2. Pharmaceutical regulations in Nordic countries

The pharmaceutical market is characterized by numerically low price elasticities on the demand side and market power on the supply side.⁴ An unregulated market would in this situation be likely to yield high pharmaceutical prices and correspondingly high expenditures of drug consumption. Most countries therefore use several regulatory instruments to con-

⁴ See Scherer (2000) for an overview of specific features of the pharmaceutical market. Brekke (2009) also offers a similar overview with a focus on the Norwegian market.

trol prices and total consumption of prescription drugs.⁵ In this section, we describe some of the most important regulations, and discuss briefly how they are expected to affect prices and the demand for pharmaceuticals. We also categorize the Nordic countries with respect to the regulatory instruments used.

2.1 Regulatory instruments

We can make a fundamental distinction between supply-side and demand-side regulation. Supply-side regulation attempts to directly control drug prices and can apply to different levels of the vertical supply chain; manufacturers, wholesalers and retailers (pharmacies). Demand-side regulation attempts to control prices and the consumption of pharmaceuticals indirectly through the design of the reimbursement system. Therefore, we can distinguish between the regulation of the price (or the margin) that the suppliers of drugs receive (supply-side regulation), and the regulation of the price that consumers actually pay (demand-side regulation).

Demand-side regulations

Health insurance implies that patients (potentially also doctors) are not very responsive to prices of alternative drugs. Insurers therefore usually do not offer a 100 percent coverage of medical expenses, but impose cost sharing on patients through copayments. The structure of the copayments is a key issue for making patients more conscious of the pharmaceutical costs. In this section, we first discuss the more regular copayment schemes that are used, and then describe a more recent and increasingly popular copayment scheme called reference pricing. Finally, we will discuss a couple of other (non-price) instruments that are employed by insurers.

The regular copayment schemes usually take two different forms: deductibles or coinsurance. Formally, we may write these two copayment schemes as follows:

⁵ Danzon (1997) offers an overview of pharmaceutical price regulations with examples from various countries.

$$c = \begin{cases} d & \text{if deductible scheme} \\ a \times p & \text{if coinsurance scheme} \end{cases}$$

where $c \in (0, p)$ is the copayment, $d \in (0, p)$ is the deductible, $a \in (0, 1)$ is the coinsurance rate, and $p > 0$ is the price of the prescribed drug. In the case of $c = 0$, there is full insurance coverage, whereas if $c = p$ there is no insurance coverage.⁶ A deductible is simply a flat fee (say EUR 5) that patients have to pay when purchasing the prescribed drug. Deductibles impose demand-side cost sharing, but the cost-sharing is regressive in the sense that more expensive drugs face a higher coverage than cheaper drugs.

A disadvantage of deductibles is that the copayment is not linked to the price of the drug. This implies that patients (or prescribing doctors) would not be responsive to the relative prices of alternative drugs. Thus, demand is likely to be price inelastic under a deductible scheme, enabling the pharmaceutical firms to charge high prices. Some insurers (countries) have therefore introduced more refined schemes with higher (lower) deductibles for expensive (cheaper) drugs. However, the correspondence between the price of the drug and the price patients face (the copayment) is still weak, so demand is not likely to be very price elastic even under multi-tiered deductible schemes.

Many insurers (countries) have therefore adopted *coinsurance* schemes, which introduce a direct link between drug prices and copayments. Under this scheme, patients pay a defined share (say 30 percent) of the price of the drug. Copayments of alternative drugs would reflect the price differences only adjusted by the coinsurance rate. Coinsurance is therefore likely to induce more price responsiveness on the demand-side and, in turn, some degree of price competition between alternative therapeutic drugs.

One issue with coinsurance schemes is that the copayments for expensive drugs can be considerable. Some insurers (countries) therefore offer a higher coverage for costly drug therapies. This can be done in several ways. One way is to impose lower coinsurance rates for more expensive drugs. Another way is to impose caps on the copayments and then offer a 100 percent coverage for additional expenditures. The disadvantage of

⁶ The latter applies to over-the-counter drugs, as well as prescription drugs that are not on the insurer's reimbursement list.

these adjustments is, of course, that they make demand less price elastic and therefore counteract the intention of coinsurance regimes.

If cost-sharing were the main concern for the insurers, then (a refined) deductible scheme could do equally well as coinsurance. However, insurers tend to prefer coinsurance to deductibles, because of the direct link to the drug price that makes demand more price elastic and increases the potential for price competition between alternative pharmaceuticals.

Let us illustrate the potential effects of coinsurance schemes on the pricing and cost-sharing between patients and insurers. In the below table, we have constructed an example with an increase in the coinsurance rate from 0.2 to 0.3. This is assumed to make demand more price elastic and trigger a price reduction by the pharmaceutical companies. In Case A the response is weak and the price is only reduced by EUR 1, whereas in Case B, the response is strong and results in a price reduction of EUR 4.

Table 2. Copayments and coverage when price responses to coinsurance rates

	Coinsurance rate	Drug price	Copayment	3 rd -party payment
Case A:	0.2	€ 10	€ 2.0	€ 8.0
Weak price response	0.3	€ 9	€ 2.7	€ 6.3
Case B:	0.2	€ 10	€ 2.0	€ 8.0
Strong price response	0.3	€ 6	€ 1.8	€ 4.2

Source: Own construction.

In Case A with a weak price response, a higher coinsurance rate is mainly shifting the costs from the insurer (third-party payer) to the patient. Increasing the coinsurance rate is therefore almost equivalent to offering a lower insurance coverage. However, in Case B with the strong price reduction, a higher coinsurance rate does not only reduce the payment for the insurer, but in fact also for the patient. In this case, a higher coinsurance rate is actually *increasing* the insurance coverage to the patients, since the *de facto* copayment has become lower. This example illustrates the two effects of coinsurance: (i) the direct effect is to shift costs from payer to patient; (ii) the indirect effect is to lower the prices of pharmaceuticals and thus, the total payment for pharmaceuticals.

Reference pricing, sometimes also called internal referencing, is a co-payment scheme that has become increasingly popular in recent years. This scheme introduces high-powered incentives for patients to choose cheaper alternative medicines. Under reference pricing, drugs are classified into different reference groups based on therapeutic effect. For each reference group, the regulator sets a reference price, which is the maximum reimbursable price for all drugs in the reference group. Any positive difference between the actual drug price and the reference price is not reimbursable. Formally, we can write the copayment under reference pricing (with coinsurance) as follows:

$$c = \begin{cases} a \times r + (p - r) & \text{if } p > r \\ a \times p & \text{if } p \leq r \end{cases}$$

where $r \in (0, p)$ is the reference price. The effect of reference pricing is to increase the price elasticity of demand for drugs priced above the reference price. The lower the reference price is set, the more price elastic demand is likely to be. Under this scheme, the insurance coverage is lower for expensive drugs, which is in contrast to deductible and coinsurance schemes. The aim of reference pricing is to induce consumers to select cheaper alternatives and stimulate price competition between producers of therapeutically related drugs.

The reference pricing schemes vary according to (i) how broadly the reference groups are defined, and (ii) how the reference price is determined. The most narrow, but also most common, definition of reference groups only includes therapeutically equivalent drugs (i.e., same substance) for which the patent protection has expired. This scheme, often called *generic reference pricing*, has the aim of inducing patients to select cheaper generic versions instead of high-priced brand-names. A less narrow definition is also based on therapeutically equivalent drugs, but extends the scheme to also including patent-protected drugs. This scheme aims at stimulating competition from parallel-imported drugs in the on-patent market segment. The more broadly defined reference pricing schemes include therapeutically related drugs (with different substances) in the reference groups. The intention of *therapeutic reference pricing* schemes is to stimulate competition from therapeutic substitutes. However, the therapeutic reference pricing schemes are also likely to limit the

profits (patent rent) of the patent-protected drugs, and are therefore more controversial from a policy perspective.⁷

Finally, the reference pricing schemes vary according to how the reference price is defined. Generally, the reference price is set somewhere between the highest priced and the lowest priced drug in the reference group. The strict regimes define the reference price equal to the cheapest drugs, implying that the patient faces a surcharge on every other drug in the reference group. In most regimes, the reference price is updated over time according to price changes by the pharmaceutical producers, and is therefore endogenously determined by market prices.⁸

There are also non-price demand-side instruments that affect the pricing and consumption of drugs. First, most insurers (countries) require the pharmaceutical firms to report cost-efficiency or a cost-benefit analysis before placing the drug on the reimbursement list. These analyses would include a suggested price by the pharmaceutical companies. Obviously, suggesting a very high price would imply a low cost-efficiency ratio, and therefore a lower probability for reimbursement. Thus, there is an implicit trade-off for the pharmaceutical companies in their price setting between a lower margin and a higher probability of getting on the reimbursement list. Some insurers use the reimbursement listing procedure actively as a negotiation tool, and exclude drugs that do not have a favorable pricing relative to the existing therapeutic alternatives.

Second, the allocation of physician's budgets for prescription drugs is an instrument that some insurers (countries) have implemented. This instrument has been used in the UK and Germany. The insurer computes a budget for each physician based on her list of patients and the cost of drugs. If the physician only prescribes high-cost drugs, the budget will quickly be spent and the patient would need to go to another physician to obtain her drug. The idea is that these budgets should induce the physicians to take into account the cost of drugs, and prescribe cheaper alternatives (e.g., generics) when available.

⁷ Brekke et al. (2007) study theoretically the effects of different reference pricing schemes, and find that therapeutic reference pricing induces stronger price competition and lower profits than generic reference pricing (or regular coinsurance).

⁸ Brekke, Holmås and Straume (2011) set up a model with endogenous and exogenous reference pricing and show that endogenous reference pricing gives generic firms a strategic incentive to lower their prices, not just to capture market shares from brand-names, but also to manipulate the reference price and make the brand-name more costly for patients.

Supply-side regulations

The supply side in pharmaceutical markets consists of a set of vertically related providers. Upstream we have the pharmaceuticals companies. These firms can be divided into two groups; brand-name and generic producers. The brand-name producers are typically innovating firms that invest in R&D and marketing, whereas the generic producers copy the original drugs and once the patent protection has expired, may enter the market with these copy products. Downstream, there are distributors (wholesalers and parallel traders) and retailers (pharmacies). There is a wide set of supply-side regulations that restrict the behavior and trade of the vertically related firms. Here, we focus on the restrictions that are aimed at affecting the pricing and demand for pharmaceuticals. This includes regulations of prices, margins, and entry into national markets.

Many insurers (countries) directly control the pricing of drugs. The most common way of controlling prices is to impose a price cap that defines the maximum price a provider can charge for a specific drug on the market. Price cap regulation obviously curbs the market power of pharmaceutical firms, but could be harmful to innovation as the profit is reduced. The interesting question is therefore how the price cap is set by the insurer.

An increasingly popular price cap scheme is *international reference pricing* (external referencing). Under this scheme, the price cap for a given drug is determined by the prices of the same drug in a set of reference countries. The exact formula for the price cap varies from country to country, but is usually a weighted average of the prices of the drug in the foreign countries. The strictness of the price cap scheme would therefore depend on the countries selected in the reference group, and whether the formula imposes a price cap at the lower end of the price distribution in the foreign countries.

International reference pricing is a simple procedure for fixing the price cap and ensures that the price level in a given country is not at the higher end. However, this scheme relies on foreign countries setting drug prices that offer optimal returns on the R&D investments. The most obvious effect of international reference pricing is that it contributes to an international harmonization of drug prices. The more countries that apply this instrument, the stronger is the effect. This scheme would therefore prevent international price discrimination by the pharmaceutical firms.

The incentives for innovation are not likely to be optimal under international reference pricing.

Price cap regulation is usually imposed at either the manufacturer or the wholesale level. To make the price cap binding at the retail level, most insurers (countries) impose a mark-up regulation on the downstream firms. One interesting issue is that different mark-up schemes could affect the final consumer prices through the pharmacies' dispensing incentives. More specifically, if pharmacy mark-ups are set as a percentage add-on to wholesale prices, pharmacies would have a financial incentive to increase their (absolute) mark-up by dispensing more expensive drugs. This incentive can be eliminated by setting the mark-up as a flat fee, implying that the pharmacies would be indifferent between dispensing a cheap or an expensive drug profitwise. However, a regressive mark-up scheme, where for instance the percentage mark-up is lower for more expensive drugs, gives incentives for pharmacies to dispense cheaper rather than expensive drugs. As we will see below, all these alternatives are currently in use in the Nordic countries.

There are also non-price instruments on the supply-side that are likely to affect the pricing and consumption of pharmaceuticals. *Generic substitution* regulation allows or requires pharmacies to substitute a prescribed brand-name drug with a cheaper generic version. This regulation is often combined with reference pricing to facilitate the sales of generics. However, the pharmacies' incentives for generic substitution depend on their financial gains from this costly activity. If the mark-up regulation is progressive (e.g., the percentage add-on on the wholesale price), then pharmacies will benefit from dispensing the prescribed high-priced brand-name. Thus, for generic substitution regulation to be effective, a regressive mark-up regulation would most likely be needed.⁹

2.3 Regulatory schemes in Nordic countries

Let us now consider the regulatory schemes in the Nordic countries according to the different instruments used in demand-side and supply-side regulation. When making this classification, it is important to bear in

⁹ Brekke et al. (2012) study the pharmacies' incentive to substitute generics for brand-names, and show that this relies on the relative product margins and copayments. Using Norwegian register data, they find that a higher margin on generics relative to brand-names is associated with a higher generic market share.

mind that many real-world regulatory schemes combine elements from the more stylized regulatory models presented above. We start by describing the demand-side regulations. Table 3 classifies the various instruments used to affect the demand in the Nordic countries.

Table 3. Demand side regulations in Nordic countries

Country	Reference pricing	Reference pricing applies to	Coinsurance	Regressive coverage
Denmark	Yes	Same substance	Yes	Yes
Finland	Yes	Same substance, off-patent only	Yes	Yes
Iceland	Yes	Same substance	Yes	Yes
Norway	Yes	Same substance, off-patent only	Yes	Yes
Sweden	Yes	Same substance	Yes	Yes

Source: Brekke et al. (2011b).

The Nordic countries make use of reference pricing schemes to limit the reimbursement and induce patients to choose cheaper versions of drugs with the same chemical ingredient. In Norway, this is not the official name given to the scheme.¹⁰ The system nevertheless has the fundamental ingredients of a reference pricing system. The same argument applies to Sweden, which does not officially use generic reference pricing. However, since it is compulsory for pharmacies to perform generic substitution, unless the patient chooses to pay the price difference between the brand-name drug and the cheapest available generic drug, the system is a *de facto* generic reference pricing scheme.

The reference pricing schemes in Denmark and Sweden are more extensive. In Norway and Finland, this scheme only applies to substances where the patent has expired and generic products have been introduced. However, in Denmark and Sweden, the reference pricing scheme also applies to patent protected products when parallel imported drugs with the same substance are introduced. In this sense, the Danish and Swedish schemes do not only exploit generic competition, but also competition from parallel trade in the on-patent market segment.

¹⁰ The scheme in Norway is called «Trinnpris», and implies a step-wise cut in the reference price (trinnpris) over time after generic entry.

Another difference between the Nordic countries is the formula for the reference price. Denmark and Sweden practice a strict scheme where the reference price is set equal to the lowest price of the drugs in a given reference group. In Norway, the reference price is a fixed discount on the brand-name price when generic entry took place. The Danish and Swedish reference prices are updated frequently (every 14 days) and endogenously determined by the price setting of the pharmaceutical firms. The Norwegian reference price is, however, exogenous and not dependent on the price setting by the firms after being exposed to reference pricing.

In addition to reference pricing, all Nordic countries have copayments based on coinsurance, but there are some significant differences. In Norway, the coinsurance rate is 38 percent. However, this is combined with copayment caps both per prescription and per year. The yearly cap also includes copayments on other health care services such as physician visits, etc. For medical expenses exceeding the cap, there is 100 percent coverage. Thus, the *de facto* cost sharing is much lower than 38 percent. Notably, the surcharges under reference pricing are not subject to the copayment caps, and have to be paid out-of-pocket irrespective of the cap. In Sweden, the coinsurance rates vary according to the price of the drug. Expensive drugs face a lower coinsurance rate than cheaper drugs. This scheme is similar to the copayment cap scheme in Norway, but less discrete in its nature.

Regarding the use of supply-side regulation, Table 4 summarizes the instruments used in the Nordic countries.

Table 4. Supply side regulation in Nordic countries

Country	Price Cap regulation	Mark-up regulation	
		Wholesalers	Pharmacies
Denmark	No	No direct regulation	Linear (% + flat fee)
Finland	Yes	No direct regulation	Regressive (% + flat fee)
Iceland	Yes	No direct regulation	Regressive (% + flat fee)
Norway	Yes	No direct regulation	Regressive (% + flat fee)
Sweden	No	No direct regulation	Regressive (% + flat fee)

Source: Brekke et al. (2011b).

First, we see that Denmark and Sweden allow free price setting on pharmaceuticals, while Finland, Iceland and Norway resort to direct price control. The *price cap scheme* in Finland and Norway is based on interna-

tional reference pricing. Norway uses a basket of nine European countries as a benchmark.¹¹ Finland uses a much wider set of countries, and includes most countries in the EEA. The price cap in Norway is fixed at the average of the three lowest prices in the reference countries. The Finnish price cap formula is less transparent, and it is based more on a “reasonable” price relative to the reference countries. Thus, it is less clear whether the Finnish system is a strict price cap regime. Iceland bases the price caps on the average Nordic prices. A specific feature of the price cap regulation in Iceland is that separate caps are set for the original product and generic drugs. Denmark and Sweden do not control prices through price cap regulation, but rely more on their extensive reference pricing scheme to stimulate price competition from parallel trade and generic producers. However, they have some degree of price negotiations when it comes to the inclusion of drugs on the reimbursement list due to the requirements related to cost-effectiveness.

The table shows that all Nordic countries practice *mark-up regulation* at the downstream level. The mark-up regulation is imposed at the pharmacy level, leaving the wholesaler margins unregulated. The regulated mark-up is based on the wholesale prices (pharmacy purchasing prices) and consists of two parts; (i) a percentage add-on and (ii) a flat fee. In Denmark, the percentage mark-up is linear (8.6 percent) irrespective of the price level. In the rest of the Nordic countries, the percentage mark-up that pharmacies are allowed to add is lower for expensive drugs, and therefore regressive. However, all Nordic countries allow the pharmacies to add a flat fee for each pack sold. Interestingly, Sweden offers a higher flat fee (SEK 10) on generics and parallel-imported drugs than on brand-names, yielding the pharmacies a financial incentive for generic substitution. In Norway, the regulation of pharmacy margins is not very effective, since more than 80 percent of the pharmacies are owned by the wholesalers. The main purpose of the mark-up regulation is therefore to set the maximum price (price cap) at the retail level.

Finally, we would like to mention that the taxation of pharmaceuticals varies across the Nordic countries. Some countries indirectly subsidize consumption of pharmaceuticals by charging a lower value-added tax (VAT) rate than on other products or services. The table below shows the

¹¹ These countries are Belgium, Denmark, Finland, Germany, Ireland, the Netherlands, Norway, Sweden and United Kingdom.

regular VAT and the VAT imposed on pharmaceuticals in the different Nordic countries.

Table 5. Value added tax rates in Nordic countries, 2011

	Regular VAT %	VAT % on pharmaceuticals	
		Prescription drugs	Non-Prescription drugs
Denmark	25.0	25.0	25.0
Finland	23.0	9.0	9.0
Iceland	25.5	25.5	25.5
Norway	25.0	25.0	25.0
Sweden	25.0	0.0	25.0

Source: EFPIA/EU (2011).

The table shows that the VAT rates vary across the Nordic countries. The governments in Denmark and Norway impose the regular VAT of 25 percent on pharmaceuticals. However, Finland and Sweden subsidize pharmaceutical consumption by charging a lower VAT than the regular one. In Finland, there is a 9 percent VAT on both prescription and over-the-counter drugs. There is no VAT on prescription drugs in Sweden, but OTC drugs are charged the regular VAT of 25 percent.

To sum up, there is a considerable variation in the regulatory schemes in the Nordic countries, perhaps surprisingly large, despite the similarities between countries in this region.

3. Do economic incentives matter?

In this chapter, we will discuss how regulatory schemes and economic incentives matter in the choice of pharmaceuticals. We first look at generic substitution, and then therapeutic substitution, based on a recent study by Dalen, Locatelli, Sorisio and Strøm (2011).

3.1 Generic substitution

From March 2001, Norwegian pharmacies were allowed to substitute a branded drug for a generic version, independent of the product name prescribed by the doctor. Being permitted to intervene between the physician and the patient, the pharmacies got an active role in the market for

generics. The doctor can still guard against substitution, but this requires an explicit reservation to be added to the prescription note (“active substitution method”).¹² If the doctor refuses to substitute on behalf of a patient who is covered by the social insurance scheme, the brand-name price mark-up (as compared to the cheapest generic version) is paid by the social insurance scheme. Even without such a reservation by the physician, the patient may insist on the branded drug, in which case the pharmacy is obligated to hand out the brand-name drug. In this case, the insurance scheme does not cover the price difference between the branded drug and the reference price. The difference has to be paid by the patient himself.

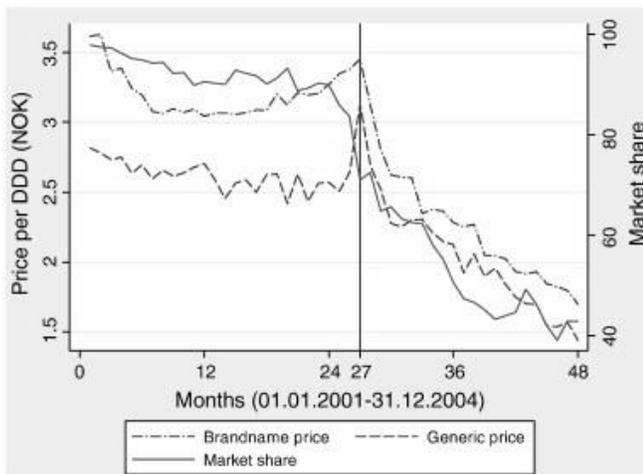
Price comparisons with other Nordic countries showed, however, that the generic substitution introduced in 2001 was not sufficient to trigger price competition and lower retail prices. The weak price response of generic substitution motivated a new regulatory scheme – “index pricing” – introduced in March 2003. The index price scheme was established for six different drugs: omeprazol (ulcer), enalapril and lisinopril (high blood pressure and heart failure), citalopram (depression), cetirizin and loratadin (allergy). Simvastatin (high cholesterol) was added in June 2004.

For these drugs, the regulator set a reimbursement price (the index price) to be paid to the expediting pharmacy, irrespective of what the chain paid for the chosen drug. This gives the pharmacies strong incentives to facilitate fierce price competition between producers of generic drugs. The index price on a drug (chemical substance) was updated every third month, and set equal to the sales-weighted average of all prices reported by the pharmacy chains, plus a fixed distribution (wholesale and retail) margin. If a retailer selected a producer with a price exceeding the average of the sales-weighted average of all prices, the net margin of the integrated retailer-wholesale pharmacy firm drops below the fixed distribution margin, whereas a retailer selecting a producer with a lower producer price experiences an increase in his net margin. This way of regularly updating the index price, based on observed producer prices from previous months, ensured that the index price tracked the development in producer prices over time.

¹² Another doctoral procedure would be the “two-line method”. Here the doctor signs either on a line that reads “brand-name necessary” or on a line that reads “substitutions allowed”. Both methods have been in use in the US, and prove to have an impact on the number of refusals. The two-line method generates more refusals than the active substitution method (Hellerstein, 1998).

The index price scheme was expected to stimulate generic substitution in pharmacies, thereby triggering price competition between producers. Brekke et al. (2009) investigated to what extent the index price scheme was successful in stimulating price competition compared to the price cap regulation. They found that index price regulation significantly reduced both brand-name (by 18-19 percent) and generic prices (by 7-8 percent). Figure 4 shows the average price of brand-name drugs and generics from January 2001 to December 2004. The vertical line indicates the introduction of the index price scheme. The price drop following the index price scheme goes together with a substantial drop in the market share of brand-name drugs.

Figure 4. Price per DDD for brand-name and generic, and brand-name market share in Norway



Source: Brekke et al. (2009).

In a follow up paper, Brekke et al. (2011a) showed that the index price regulation also triggered a significant shift in market shares towards generic drugs, which together with the price reductions resulted in substantial cost savings of about 30 percent.

In January 2005, the index price scheme was replaced with a new price regulation scheme that abandoned the direct use of economic incentives to bring down pharmaceutical prices after patent expiration. The new scheme – called the *step-wise* (no. “Trinnpris”) model – consists of a predefined, stepwise reduction of the reimbursed price, starting from the

time of generic entry into the market. The pharmacies are instructed to have the drug available at the reimbursable retail price.

The scheme gives the pharmacy chains strong incentives to lower their purchasing prices. The model does not prescribe any future price reviews based on the development of these prices. All cost savings – in terms of reduced purchasing prices – are kept by the pharmacies themselves. This scheme illustrates the fundamental trade-off that often has to be made in the regulation of prices. Maximum incentives to minimize costs (here to put pressure on the producer prices of generic drugs) are obtained by offering fixed retail prices. However, in order to be credible, these prices must be set at sufficiently conservative levels. If the government is too eager in reducing the cost of drug reimbursement – by setting the post-generic prices at very low levels – the pharmacies will report economic problems which, in turn, will make it necessary for the government to increase the prices. When such a scheme could be enforced without protests from the pharmacy chains, there are good reasons to expect the pre-determined prices to be pleasantly higher than the purchasing prices.¹³

Dalen, Furu, Locatelli and Strøm (2011) investigate how prices, different regulatory schemes, and characteristics of patients, doctors and pharmacies affect the substitution between brand-name and generics. The choices of patients/doctors and pharmacies are modeled as a bilateral comparison between the utilities of using brand-name and generics.

The analysis is not able to disentangle the doctors' prescriptions and the choices made by the pharmacies, given that the doctor has not blocked for a generic substitute. The choice probabilities in the model are thus the product of the prescription probability and the choice probability of the pharmacies. In the empirical model, unobserved heterogeneity in the choice probabilities is allowed for.

Data used in the estimation of the model was extracted from the Norwegian Prescription Database (NorPD) at the Norwegian Institute of Public Health. NorPD was established on January 2004. All drugs in Norway are classified according to the Anatomical Therapeutic Chemical (ATC) classification system. From this database, Dalen, Furu, Locatelli and Strøm (2011) extracted the entire population of prescriptions in February 2004 and 2006 for 23 different drugs (chemical substances) subjected to generic competition. This amounts to 313 078 observations (102 201 in

¹³ The step-wise model was proposed by the pharmacy chains.

February 2004 and 210 877 in February 2006). Between 2004 and 2006, several drugs were opened for generic entry, and this explains the increase in the number of observations. The reason for adding February 2004 as well as February 2006 was to capture the two regulatory schemes: the “index price” and “the step-wise price” model.

With up to 23 chemical substances, they are able to cover a broad set of indications, such as blood pressure and heart failure, cholesterol, depression, ulcer, antibiotics, and allergy. Several of the drugs in the study are among the most selling drugs in Norway, which was also the motivation for selecting these drugs. The drugs include simvastatin (cholesterol), cetirizin (allergy), and enalapril (blood pressure).

Their empirical results imply that the larger the difference is between the price of the brand-name and generics, the less likely it is that the brand-name is purchased. Thus, generic substitution works.

Patients with prescriptions covered by the national insurance scheme (No. “Blåresept”) are more likely to use the brand-name drug. Moreover, older patients are less likely to end up with a generic and older doctors are more likely to prescribe brand-names. They also find that time after generic entry matters: The probability of generic prescription increases with time after generic entry – generic substitution is less likely for young off-patent molecules.

Of particular interest is the result of the impact of the index price scheme on the choice of brand versus generics. As mentioned above, in 2004 this scheme should give the pharmacy an incentive to dispense cheaper versions. In Dalen, Furu, Locatelli and Strøm(2011), there are four chemical substances that were covered by this scheme in 2004. For these substances, the probability of choosing brand-names turns out to be lower than for other substances. The impact is strong, with a 26 percent lower probability of choosing brand-name versions instead of a generic drug.

This result is in line with the results derived by Brekke et al. (2009). As mentioned above, they find that the index-price scheme had a significant and strong impact on prices, both for generic and brand-name versions. Note that although the prices dropped and more so for the brand-name, the price of the brand-name was still higher than for the generics.

Another study of generic substitution in the Norwegian market is Brekke et al. (2012). This paper uses register data to compute the gross

margins that the pharmacy chains have on selling brand-names and generics. The study reveals that the pharmacy chains obtain higher margins on generics, and find a strong, positive relationship between relative margins and the products' market shares. They also show that this effect is stronger for the products under reference pricing. The results indicate that the pharmacies are more likely to promote a generic substitute to patients, the larger is the generic margin relative to the brand-name margin. Thus, financial incentives are important for pharmacies' incentives to engage in generic substitution.

Sweden introduced generic substitution in October 2002. Pharmacies were required to substitute the cheapest available generic for the brand-name prescribed by the doctor. As in Norway, patient copayment increased if the cheapest drug was not chosen. Granlund (2010) investigates the effect of the reform on prices and demand using panel data from 1997 to 2007. He finds that the introduction of generic substitution on average lowered the prices by 10 percent. The price drop was strongest for brand-name drugs that faced generic competition prior to the reform. For these drugs – the price-drop was 14 percent.

A Finnish study reveals similar price responses to generic substitutions, introduced in March 2003. Aalto-Setälä (2008) finds that the reform led to an average price drop of 10 percent.

3.2 Therapeutic substitution: The market for Tumor Necrosis Factor (TNF) alpha inhibitors¹⁴

Using a unique natural policy experiment in Norway, Dalen, Locatelli, Sorisio and Strøm (2011b) have investigated to what extent the price responsiveness of prescription choices is affected when the identity of the third-party payer changes and the choices are made between different drugs developed for the same diagnoses (therapeutic competition). The case in point is the Norwegian market for Tumor Necrosis Factor (TNF) alpha inhibitors.

When the market for TNF-inhibitors opened in Norway in 2000, the first entrant *Enbrel* was fully covered by the obligatory national insurance plan. Treatment with *Enbrel* is initiated by the hospital doctor, but the

¹⁴ Dating TNF-alpha inhibitors representing the most important way of treating arthritis and other autoimmune diseases (Feldmann and Maini, 2003).

cost was automatically covered by the national insurance plan. The second entrant *Remicade* did not obtain the same type of coverage. Instead, the treatment cost had to be fully covered by the doctor's affiliated hospital. Importantly, the hospitals' budget did not include any earmarked grants for these patients. The cost of treatment with *Remicade*, therefore, competed with other expenses within the hospital. This sharp asymmetry in funding schemes reflects a quality attribute of the two drugs. *Enbrel* is administrated by the patients themselves (pump injections), while *Remicade* requires several hours' infusion at hospitals. In the fall of 2002, the government modified the plan for *Remicade*. The government required a copayment of 20 percent from the doctor's affiliated hospital. *Enbrel* maintained its full insurance plan coverage. The third entrant *Humira* is also administrated by pump injections by patients, and received the same funding plan as *Enbrel* when the drug entered in January 2003.

An important policy change took place in 2006. Then, the asymmetry of financing among *Enbrel*, *Humira*, and *Remicade* was entirely removed by returning the entire funding responsibility to the hospitals for all three drugs. Since then, all costs of treatment with TNF-alpha inhibitors have to be covered by the doctors' affiliated hospital.

When estimating how economic incentives affect the choices of medical doctors and patients, one has to take into account that it is not only economic incentives that matter for the choices. The quality of the pharmaceuticals as well as side effects may have an impact on the choices. It is then important to consider the obvious fact that these quality and side effects may be priced out in the market by the producers of the drugs.

To deal with this problem, Dalen, Locatelli, Sorisio and Strøm (2011) jointly estimated the market share and price-setting equations, assuming monopolistic competition. The estimated coefficients imply that doctors appear to be significantly more price-responsive when the costs are covered by the hospitals as compared to when the costs are covered by national insurance.

As expected, the numerical values of the own price elasticities increase when quality aspects are accounted for (demand and price setting model) compared to when they are not (the market share model). The mean value of the own-price elasticities is given in the table below.

Table 6. Mean own-price elasticities

	The market share model	Demand and price setting approach
Enbrel	-0.59	-2.19
Humira	-0.92	-3.39
Remicade	-0.29	-1.02

Source: Dalen, Locatelli, Sorisio and Strøm (2011).

Thus, when quality and side effects are accounted for in the model, the numerical values of the own-price elasticities become much higher. In this case, therapeutic substitution between patented drugs therefore seems to be highly price-responsive.

4. Should pharmaceutical costs be curbed?

We end this paper with a discussion of the problem raised by the title of our article – should pharmaceutical costs be curbed? Especially if we look at the Norwegian case, we see that pharmaceutical expenditure stopped increasing around 2005. Since then, we have even seen a slight drop in expenditure. This is not caused by a drop in the volume of pharmaceuticals taken by Norwegian patients. The number of defined daily doses has been increasing after 2005.

The main explanation for this combination of reduced expenditure levels and increased consumption of medicines is the increasing number of drugs that went off patent. For many years, The Association of the Pharmaceutical Industry in Norway (LMI) reported the market share of innovative drugs in Norway in their annual report “Facts and figures”.¹⁵ During the 1990’s, innovative drugs represented an increasing part of the total sales volume. Since then, however, the innovation rate has declined with few new drugs and an increasing number of drugs that went off patent. In 2000, the market share of innovative drugs was reported to be close to 38 percent. In 2005, the market share was as low as 10 percent. The flattening expenditure curve in Norway is therefore explained by a more mature pharmaceutical market in combination with a regulatory

¹⁵ Innovative drugs are here defined as drugs that have entered the market during the last five years.

policy that has enabled a strong price competition on off-patent drugs (generics). Lowering the pharmaceutical costs by implementing fierce generic competition is welfare-improving and comes without any severe negative side-effects.

However, it is of more interest to return to the political debate in the early 1990's. During these years, pharmaceuticals costs were steadily increasing and caused increasing costs for the social insurance scheme. The government repeatedly expressed concerns for the expenditure growth, which was higher than the overall growth rate in health care cost.

Since the growth rate was to a large extent caused by the introduction of new drugs, it is less clear if this is something to curb. New drugs improve treatment which is to the benefit of patients, and these benefits should be compared with the costs of funding pharmaceuticals.

In a series of papers, Frank Lichtenberg has empirically investigated the health effects of new drugs. Lichtenberg (2012a) shows that about one-third of the increase in German life-expectancy during 2001-2007 can be explained by the replacement of older drugs with newer drugs. Lichtenberg (2012b) investigates the effect of new drugs on functional limitations of elderly Americans in nursing. Functional limitations are significantly lowered by the use of newer drugs at nursing homes.

References

- Aalto-Setälä, V. (2008), The impact of generic substitution on price competition in Finland, *European Journal of Health Economics* 9, 185-191.
- Almås, I. (2012), International income inequality: Measuring PPP bias by estimating Engel curves for food, *American Economic Review* 102, 1093-1117.
- Arnberg, K., Linner, L. and Lunding, D. (2012), *Prisutveckling på läkemedelsområdet i ett internationellt perspektiv: En internationell prisjämförelse av läkemedel utan generisk konkurrens*, Tandvårds- och Läkemedelsförmånsverket, Stockholm.
- Brekke, K.R. (2009), *Markedet for legemidler: Regulering, konkurranse og utgifter*, in K. Haug, O.M. Kaarbøe and T. Olsen (eds.), *Et helsevesen uten grenser?*, Cappelen Akademiske Forlag, Oslo.
- Brekke, K.R., Grasdahl, A. and Holmås, T.H. (2009), Regulation and pricing of pharmaceuticals: Reference pricing or price cap regulation?, *European Economic Review* 53, 170-185.
- Brekke, K.R. and Holmås, T.H. (2012), *Prices of pharmaceuticals: A comparison of prescription drug prices in Sweden with nine European countries*, Report 01/12, Institute of Research in Economics and Business Administration, Bergen.

- Brekke, K.R., Holmås, T.H. and Straume, O.R. (2011a), Reference pricing, competition, and pharmaceutical expenditures: Theory and evidence from a natural experiment, *Journal of Public Economics* 95, 624-638.
- Brekke, K.R., Holmås, T.H. and Straume, O.R. (2011b), Comparing pharmaceutical prices in Europe. A comparison of prescription drug prices in Norway with nine western European countries, Report 11/11, Institute of Research in Economics and Business Administration, Bergen.
- Brekke, K.R., Holmås, T.H. and Straume, O.R. (2012), Margins and market shares: Pharmacy incentives for generic substitution, NHH Discussion Paper, Bergen.
- Brekke, K.R., Königbauer, I. and Straume, O.R. (2007), Reference pricing of pharmaceuticals, *Journal of Health Economics* 26, 613-642.
- Dalen, D.M, Furu, K., Locatelli, M. and Strøm, S. (2011), Generic substitution. Micro evidence from register data in Norway, *European Journal of Health Economics* 12, 49-59.
- Dalen, D.M., Locatelli, M., Sorisio, E. and Strøm, S. (2011), A probability approach to pharmaceutical demand and price setting: Does the identity of the third-party payer matter for prescribing doctors?, CES Working paper 3643, Munich.
- Danzon, P.M. (1997), *Pharmaceutical Price Regulation: National Policies versus Global Interests*, The AIE Press, Washington DC.
- Danzon, P.M. (1999), *Price Comparisons for Pharmaceuticals. A Review of U.S. and Cross-National Studies*, American Enterprise Institute for Public Policy Research, Washington DC.
- Feldmann, M. and Maini, R.N. (2003), TNF defined as a therapeutic target for rheumatoid arthritis and other autoimmune diseases, *Nature Medicine* 9, 1245-1250.
- Granlund, D. (2010), Price and welfare effects of a pharmaceutical substitution reform, *Journal of Health Economics* 29, 856-865.
- Health at a glance (2011), OECD indicators, Paris, <http://www.oecd.org/health/healthataglance>.
- Hellerstein, J. (1998), The importance of the physician in the generic versus trade name prescription decision, *RAND Journal of Economics* 29, 109-136.
- Lichtenberg, F.R. (2012a), Contribution of pharmaceutical innovation to longevity growth in Germany and France, 2001-7, *Pharmacoeconomics* 30, 197-221.
- Lichtenberg, F.R. (2012b), The effect of pharmaceutical innovation on the functional limitations of elderly Americans. Evidence from the 2004 National Nursing Home Survey, NBER Working Paper 17750.
- Scherer, F.M. (2000), The pharmaceutical industry, in A.J. Cuyler and J.P. Newhouse (eds.), *Handbook of Health Economics*, North Holland, Elsevier, Amsterdam.

Comment on Brekke, Dalen and Strøm: Should pharmaceutical costs be curbed?

Helgi Tómasson*

The paper by Brekke, Dalen and Strøm starts by reporting some facts on pharmaceutical sales in Denmark, Finland, Norway and Sweden in the year 2011. It is stated that the pharmaceutical cost per capita ranges from 310 to 400 euros. The authors mention growth of sales, and correctly state that not much can be inferred on data for only two years. The sales are decomposed into price and volume and it is stated that Norway tends to have the cheapest drugs. They correctly cite that exchanging, say, Swedish prices and Norwegian prices would affect price indices as the consumption weights differ between the two countries. The authors also discuss the impact of packaging. The drugs are packed differently among the countries which adds to the heterogeneity. The volume of the consumption is highest in Sweden but lowest in Denmark. The low consumption and high sales in Denmark are due to the high price of drugs in Denmark. All these facts support the claim that an international comparison is difficult.

The authors characterize the nature of the pharmaceutical market as a low price elasticity on the demand side and strong market power on the supply side. These oligopolistic characteristics result in widespread regularization systems around the world. The authors summarize some important regulatory instruments. The arsenal of instruments is large both on the demand- and the supply side, as well as for price-based and non-price-

* University of Iceland, helgito@hi.is.

based instruments. The Nordic countries seem to use most of these instruments. The authors review some differences in how the Nordic countries implement the instruments, e.g., how they treat on-patent drugs, what kind of mark-up is allowed, taxes, etc.

Next, the authors turn to the questions, “Do economic incentives matter?” and “What is a good policy?”. The econometric toolbox seems quite similar to what is described in Train (2009) and what the authors have used in other publications. Dalen et al. (2011) is cited and a model from this reference is described. To fully understand the econometrics in this paper, it is necessary to look up the formulas and concepts in these references. A large data set containing all Norwegian prescriptions from February 2004 to February 2006 is described. A scenario of the behaviour of doctors and patients is drafted. Even here, the paper would benefit from a formal statement of the model. Perhaps a graph showing the timing of the events could be helpful.

The authors report an interesting experiment, the *index pricing*, introduced in 2003. The idea of the index pricing strategy is to fix the reimbursement for some popular categories of drugs irrespective of what the pharmacy paid for each brand of drug.

In 2005 a new experiment, the *stepping price model*, replaced the index pricing model. The authors conclude that the regulatory instruments seem to have an impact, in particular they favour the index pricing model. The question of the impact on health seems to be left open. That question might be a much more challenging one.

The authors have used a large data set. The quality of the information of such a data set, and any data set, depends on the exact definition of the scientific model and the corresponding statistical model.

The paper is essentially a literature overview of, first, the regulatory environment of pricing of pharmaceutical products in the Nordic countries and, second, a discussion of a few econometric results that the authors have derived in earlier research (Brekke et al., 2009; Brekke et al., 2011) on responses and preferences in the pharmaceutical market. The conclusion seems to be that the pharmaceutical costs can be curbed, but it is less clear whether they should be.

References

- Brekke, K.R., Grasdahl, A.L. och Holmås, T.H. (2009), Regulation and pricing of pharmaceuticals: Reference pricing or price cap regulation?, *European Economic Review* 53, 170-185.
- Brekke, K.R., Holmås, T.H. och Straume, O.R. (2011), Reference pricing, competition, and pharmaceutical expenditures: Theory and evidence from a natural experiment, *Journal of Public Economics* 95, 624-638.
- Dalen, D., Furu, K., Locatelli, M. och Strøm, S. (2011), Generic substitution: Micro evidence from register data in Norway, *European Journal of Health Economics* 12, 49-59.
- Train, K. (2009), *Discrete Choice Methods with Simulation*, 2nd edition, Cambridge University Press, Cambridge.

Productivity differences in Nordic hospitals: Can we learn from Finland?

Clas Rehnberg* and Unto Häkkinen**

Summary

Acute short-term hospitals are the major resource user in the health care sector and play a significant role for advanced treatment. This paper presents the findings from a Nordic collaboration where the productivity differences across acute hospitals have been measured and compared. The results suggest that there was a markedly higher average hospital productivity in Finland as compared to Denmark, Norway and Sweden. The explanations of findings are discussed along different theories and possible reasons for the observed differences. The findings do not seem to be explained by differences in the use of market mechanisms and reimbursement systems. The paper argues for a closer analysis of the impact of fund-holding, contractual relations and incentives between levels of public governments as well as including quality indicators in the efficiency measure.

Keywords: productivity, hospitals, benchmarking, DEA, Nordic countries.

JEL classification numbers: C14, D23, D24, I18.

* Medical Management Centre, Karolinska Institutet, Clas.Rehnberg@ki.se.

** National Institute for Health and Welfare, Centre for Health and Social Economics, Helsinki, Finland, unto.hakkinen@stakes.fi.

The performance of health systems has become an interest for both health economists and policy-makers in the health sector (Hurst and Jee-Hughes, 2001). Comparative studies of health system performance have been published by researchers and by a number of international organizations (Arah et al., 2006; Joumard et al., 2010). Many OECD-countries have developed national performance measurement frameworks for monitoring and comparing the overall efficiency of their health care systems (Häkkinen and Joumard, 2007). Most of these studies analyze the entire health systems where the different performance measures are compared and sometimes related to the use of resources. There are relatively few analyses of hospital efficiency and productivity that are based on cross-national datasets.

This paper presents the findings from a Nordic collaboration where the productivity differences across acute hospitals have been measured and compared. The explanations of findings are discussed along different theories of relevance and possible reasons for the observed differences. The Nordic health systems have several similarities but also interesting differences. The countries express clear goals and aspirations concerning an equal and universal access to health services. There are also common structural features such as tax-based funding and an important role for regional and local governments (Magnussen et al., 2009; Iversen, 2011). The decentralized structure is exercised by local political governance. Still, there are important differences in how the countries approach such issues as governance, financing, and contracting as well as the role of patients in terms of choice and rights. The four countries also share many of the administrative tools and mechanisms for running the health systems. Some of these imply common standards for registering utilization and outcomes, which make cross-country comparisons easier to conduct.

The outline of the paper is as follows. The next section gives an overview of the health care systems in the Nordic countries with an emphasis on relevant factors explaining the productivity differences. Section 2 provides the result of the productivity analysis from the Nordic collaboration. In Section 3, the findings are discussed against some theories and structural differences across countries. Section 4 discusses what we know and what we do not know about explaining the differences in hospital productivity and gives suggestions for further research.

1. Structure of the Nordic health systems

The health care systems in Europe are often characterized as tax based systems or social health insurance systems.¹ The Nordic countries belong to the former group as health care is financed through taxes, most services are free of charge and most services are provided within the public sector. In contrast to other tax-based systems, like the National Health Service (NHS) models in the United Kingdom, New Zealand and southern Europe, the Nordic countries give a more pronounced role to local and regional governments for decisions about financing, allocation of resources and provision of services. The multilevel governing process is a key feature in the Nordic model with differences across countries. Recently, Denmark and Norway changed the decentralized structure into a more centralized model, whereas Finland and Sweden retain a higher degree of decentralization.

The Nordic model has a historical common base with a multilevel public sector. Characteristically, the financing and organizing of health care has been decentralized to regional and local governments. This decentralization applies to the political level with financing through local taxes and local provision of services (Pedersen, 2004). In economic terms, the underlying idea is that of fiscal federalism (Rattsø, 2002; Tiebout, 1956). However, there are important differences in how health care is financed and produced. Even if all four countries show a decentralized structure, there are differences when it comes to such functions as financing, regulation and provision. The health systems display a considerable diversity in terms of sources of funding, the purchasing function and the provider structure. The interactions between these functions are also organized differently. Denmark, Norway and Sweden show a similar history with regional governments (counties) for all three functions (political, financial and provision). Finland has the most far-reaching decentralization of the health care function and responsibility with the municipalities being in charge and financing most of the health care.

Public financing by taxes covers the entire population in all four countries and comprises basic as well as advanced health care. The solidarity principle through a tax-based financing on different levels has remained unchanged after the centralization of financing to the central government

¹ Sometimes labeled Beveridge or Bismarck-systems, respectively.

level in Denmark and Norway. The decentralized system of local taxes in Denmark and Norway changed at the beginning of the new millennium when both countries reorganized financing as well as provision towards a more centralized health system. After the local government reform in Denmark in the year 2007, the new regions could not finance health care by levying local income and property taxes. With the reform, the central government levies taxes for health care and resources are then allocated to the regions based on needs criteria. The purpose of the 2007 reform in Denmark was to ensure a greater equality in hospital treatment across the country, by increasing the influence of the National Board of Health on hospital planning. The number of regional authorities was reduced from 14 counties to 5 regions. The municipalities received more responsibility for rehabilitation, disease prevention and health promotion, as well as the care and treatment for disabled people, and alcohol and drug users. Municipalities contribute to the regions through payments both per capita and by activity, the latter according to citizens' utilization of the regional health services (Pedersen, 2004).

In Norway, the financing is divided between the central government and the municipalities. The four regional health authorities are funded by the central government through a combination of need-based grants and activity-based funding. Hospitals are almost entirely reimbursed by the four regions. In the early 2000's, the Finnish municipalities financed over 60 percent of hospital care from local taxes and the central government approximately 30 percent of hospital services via non earmarked state subsidies. Although the subsidies have increased quite rapidly in recent years, still about 55 percent of the hospital services are financed by local municipal taxes, where the level is decided locally. Hence, the municipalities control a major part of the health care expenditures. In Sweden, the regional county councils levy taxes and have the power and responsibility to decide on local tax rates. In addition, the central government transfers grants based on needs and differences in cost structure to the regional level. To sum up, both Denmark and Norway have centralized the control of health care expenditures, whereas Finland has a far-reaching decentralized fiscal control, and Sweden has chosen an intermediate arrangement.

The provision of most services is delivered by local and regional governments. The exception has been the primary health care sector where Denmark has a tradition of private general practitioners (GPs), a system

that Norway adopted in 2002 and Sweden is gradually moving towards. In Finland, most GPs are employed by the municipalities but an increasingly larger share of GPs is working in occupational care providing primary health service. The hospital sector is almost entirely under public ownership, although there is a variety in terms of payment mechanisms and the level of organization. The Norwegian reform in 2002 removed the responsibility from the 19 counties and transferred the ownership of hospitals to five (and later four) regions (*regionale helseforetak*), each with its own professional board. The essential idea was to create what is akin to a private corporate structure: a corporation (*regionale helseforetak*) with its own board and hospitals (*helseforetak*) as subsidiaries, also with their own boards. In Denmark, most secondary and tertiary care takes place in general hospitals owned and operated by the regions. Consequently, the reform from 2007 reduced the number of hospital owners (Pedersen, 2004). As a modern hospital requires a catchment area of a minimum size, most of the Finnish municipalities are too small to run their own acute hospitals. There is a long tradition in Finland that each municipality is obliged to be a member of a hospital district. The municipalities indirectly own the hospitals, but the financing (or reimbursement) could be considered as an *at arm's length* relationship where each municipality pays according to its use of service to a hospital district. The decentralized nature of the system and the absence of national guidance of the payment principles imply a considerable variation across hospital districts in the design of schemes. The municipalities run their own primary health care centers and nursing homes themselves or through a joint ownership with other municipalities. From 2007, the municipalities started to reorganize models of production. The purpose of the local reforms is to enhance the co-operation between primary and secondary health care and social service.² In Sweden, the regional county councils both finance and run the acute hospitals. In some county councils, the payments to the hospitals are handled through negotiation and administratively set prices in a purchaser-provider split arrangement. In a few county councils, the hospitals have been transformed into joint-stock companies, but where all stocks are owned by the county council. One private for-profit hospital is

² The reforms include the merging of health centers and regional hospitals into one organization, creating a new regional self-regulating administrative body for all municipal services. (Social services, upper secondary schools and vocational service are included in addition to health service, see Häkkinen and Jonsson, 2009).

located in Stockholm with a long-term contract of 7-9 years with the county council as the funder.

The level of co-payment is low in all Nordic countries following a policy that health services should not be rationed by the price-mechanisms. However, the types and levels of co-payment vary between the countries. The Danish system has for a long period of time worked without cost-sharing policies in the GP and the specialist sector. The other countries have a patient fee for GPs and specialist out-patient services. Neither Denmark nor Norway has co-payment for hospitalization, whereas Finland and Sweden charge low fees for inpatient care. In all countries, there is an upper limit ('ceiling') for patient yearly expenses. Sweden is the only country without a clear *gate-keeping* system, which is used by the other countries through the GPs for moderating access to hospitals.

There are considerable differences in the cost spent on health services between the Nordic countries. A commonly used indicator is health care spending as a share of the gross domestic product (GDP), which shows that Finland spent around 7-8 per cent of its GDP which is less than the other Nordic countries with a spending of around 9-10 per cent of their GDP during the period. Norway initially spent a lower share of its GDP on health, but now has the same level as its neighbors. Using another indicator as the health care spending per capita shows that Norway has the highest level of expenditures of USD 4 076 PPP³ in the year 2004. Denmark and Sweden spend around USD 3 000 PPP/capita and Finland USD 2 450 PPP/capita the same year. The difference is partly due to differences in the cost level and partly to differences in resources use. The Norwegian system has also been the most expansive system since the early 1990's, both in terms of the share of GDP spent on health and the overall growth of real spending per capita (OECD, various years). This is mainly explained by its strong economic growth, and the Norwegian spending on health as a share of GDP fits well into models of income elasticity.

To sum up, the Nordic health systems have several similarities and but also interesting differences. The countries share common objectives for a universal health system based on taxation. The role of local and regional governments has been a common feature, although recentralization reforms have been implemented in Denmark and Norway. The decentral-

³ PPP = purchasing power parity.

ized structure is maintained in Finland and Sweden but with important differences regarding how health services are financed and contracted by the local governments. Finally, on an aggregated level, Finland has for a longer period of time spent less of its GDP on health services than the neighboring countries.

2. Differences in hospital productivity – Nordic countries

This section summarizes the work on hospital efficiency carried out by the Nordic Health Comparison Study Group (NHCSG).⁴ In all studies, an effort has been made to push the limits of the international comparison of hospital efficiency further by using patient-level data from several countries. Such comparisons are rare due to differences in the measurement of input and output definitions, but are also due to how the patient case-mix is measured across countries. Examples of previous studies are Hansen and Zwanziger (1996) who used cost functions to compare marginal costs in general acute care among US and Canadian hospitals. Mobley and Magnussen (1998) examined the relative performance of Norwegian and Californian hospitals using Data Envelopment Analysis (DEA) and empirical data from 1997. Otherwise, most comparisons are in-country analysis.

The studies carried out by the research group was done in the four Nordic countries (Denmark, Finland, Norway and Sweden) in a setting where the structure of organizing hospital care and the available data (e.g. coding and used primary classifications) are sufficiently similar. In addition, each of the Nordic countries applied similar DRG⁵ grouping systems for hospital admissions based on a common Nordic NordDRG grouping system. The aims of the different studies depart from a comparative perspective of acute hospitals in the Nordic countries, but each sub-study has its own objective and different datasets in terms of time period and selection of hospitals:

⁴ The NHCSG was set up by the National Institute for Health and Welfare, Finland, Danish Institute for Health Services Research, Copenhagen, Denmark, Ragnar Frisch Centre for Economic Research, Oslo, Norway, SINTEF Health Research, Trondheim, Norway and Karolinska Institutet, Stockholm, Sweden.

⁵ Diagnosis Related Groups – a classification system for hospital cases and products.

- to compare the performance of hospital care in the Nordic countries (Linna et al., 2010) and to investigate whether the Norwegian hospital reform has improved hospital productivity using the other four major Nordic countries as controls (Kittelsen et al., 2008). Both studies are based on data 1999-2004.
- to estimate the cost efficiency of producing patient care as well as clinical education and clinical research activities at university hospitals in the Nordic countries (Medin et al., 2011), based on data for 1999-2004.
- to compare the performance of hospital care in the Nordic countries at the hospital and regional level (Kittelsen et al., 2009; Kalseth et al., 2011), based on data for 2005-2007.

The aim of this paper is to describe the findings from these studies and discuss the explanatory factors. Even if the papers have a different focus, there are some major findings pointing at similar conclusions about the overall hospital efficiency across the countries which are robust. Still, there is a need to find explanatory factors on a health system level.

2.1 Data and Method

Meaningful international cost efficiency comparisons must be based on comparable data. In the different sub-studies by the NHCSG hospital, discharge data grouped in diagnosis related groups (DRGs) were used. A common set of DRGs was defined, and weighted using information about an average of country-specific relative costs. Inputs were measured as operating expenses, exclusive of capital, and deflated using PPPs and a specially constructed wage index. Data at the hospital level were collected from Norway and Finland for the period 1999-2004, from Sweden for the period 2001-2004 and from Denmark for 2002. A total of more than 700 hospital observations were thus included in the different analyses. The university study included a dataset of 70 university hospital observations over three years. The following inputs and outputs were included in the analyses (Table 1).

Table 1. DEA-models – Inputs and outputs and number of hospitals

	All hospitals		University hospitals
Inputs	Operating costs in real value		Operating costs in real value
			Teaching and research costs
Outputs	Surgical inpatients DRGs		DRG-adjusted surgical hospital cases
	Medical inpatients DRGs		DRG-adjusted medical hospital cases
	Surgical day patients DRGs		Outpatients visits
	Medical day patients DRGs		Postgraduate medical students
	Other DRGs		Doctors under supervision
	Outpatients visits		No. of citations
			CWTS field normalised citation score [*]
			Share of top 5% publications
No. of hospitals	Denmark:	54 (2002)	10 (2002)
	Finland:	230	15 (2002-2004)
	(1999-2004)		
* years	Norway:	256	24 (2002-2004)
	(1999-2004)		
	Sweden:	188	21 (2002-2004)
	(2001-2004)		

Source: Kittelsen et al. (2008); Medin et al. (2011).

Note: ^{*} The indicator developed at the Centre for Science and Technology Studies (CWTS), Leiden University, corresponds to the relative number of citations of publications from a specific unit compared to the world average of citations of publications of the

In the study covering all Nordic hospitals, the outputs based on the NordDRG were defined in six broad output categories as shown in the table. The university study used the same data, but patient-related costs and patient care output were supplemented by costs for teaching and research, teaching outputs and research activities. Research activities were measured by the results of a bibliometric analysis in clinical medicine (all pre-clinical research was excluded).

An input price index was developed to adjust for wage and price differences. Inputs were initially measured as operating costs in each country's national currency. The input price index was based on an assumption that operating costs and teaching and research costs are distributed among three inputs: physicians (20 percent), nurses (50 percent) and other inputs, such as materials, equipment and rents (30 percent). The proportions of costs for physicians and nurses were weighted using a wage index based on country-specific official wage data, including all personnel costs (i.e. pension costs and indirect labor taxes). Other costs were adjusted with a harmonized CPI from Eurostat and converted into euros using a

purchaser power parity corrected price index from OECD. Finally, a Paasche index is constructed using Finland in 2004 as the reference point.

The most frequently used techniques for measuring cost efficiency in health care production are applications of parametric stochastic frontier (SF) methods or nonparametric data envelopment analysis as developed, in among others, Charnes et al. (1978). This paper focuses on the findings and policy implications from the Nordic studies. Methodology issues are discussed in Kittelsen et al. (2008) and Linna et al. (2010). The DEA approach is nonparametric and less prone to specification error because of milder conditions set for the form of technology. Therefore, a cost-minimization behavior, which is not the regular case in the public sector, need not be assumed. The DEA is also easier to handle in organizations with multiple outputs and inputs. In this paper, the summary of the NHCSG's work is limited to the analysis based on the nonparametric DEA.

The DEA was used in assessing the cost efficiency of hospitals which utilizes linear programming techniques in the calculation of unit-specific efficiency scores. DEA constructs a piecewise linear efficient frontier which serves as the reference in the evaluation of efficiency. If a hospital is efficient, it lies on the frontier and will receive an efficiency score of 1.0 (100 percent efficiency). Inefficient hospitals will receive a score lower than 1.0. For example, if the score for a hospital is 0.80 as measured in the input direction, its inefficiency is 20 percent and it could produce its output with 20 percent less input. Alternatively, with an output efficiency score it produces 80 percent of its potential and it could increase its output by 25 percent using the same resources.

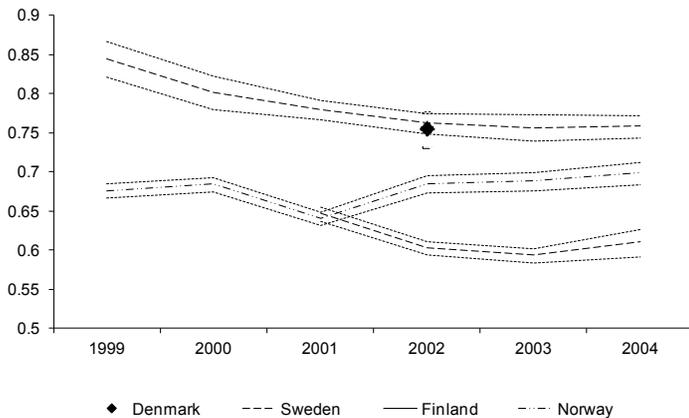
In the analysis, the bias is related to the number of observations in the sample, the number of inputs and outputs and the density of observations around the relevant segment of the frontier. The small sample bias can be remedied if knowledge of the sampling distribution is available. One method for obtaining sampling distributions of the frontier estimates is by using the bootstrap method. By calculating standard errors and confidence intervals of the indices from simulated distributions of corresponding indices in pseudo samples, bias-corrected efficiency estimates were obtained. The cost efficiency scores were calculated under the assumption of a time invariant production frontier (pooled sample). The DEA esti-

mates and bootstrap bias corrections and confidence intervals were calculated using the Frisch-DEA software package.

2.2 Results

The first analysis on hospital level data was published in Kittelsen et al. (2008) and revealed considerable differences in cost efficiency between the Nordic hospitals. The average efficiency was highest in Finland for all years (1999-2004), followed by Denmark (only year 2002) and Norway. Sweden appeared to have the least efficient hospitals. While the individual hospital scores and even the country average efficiency scores varied markedly in different model specifications, the rank of the country group averages remained the same in all models used (Linna et al., 2010). In Figure 1, the bias-corrected efficiency scores are presented for all countries.

Figure 1. Average bias-corrected productivity levels and 95% confidence intervals by country and year



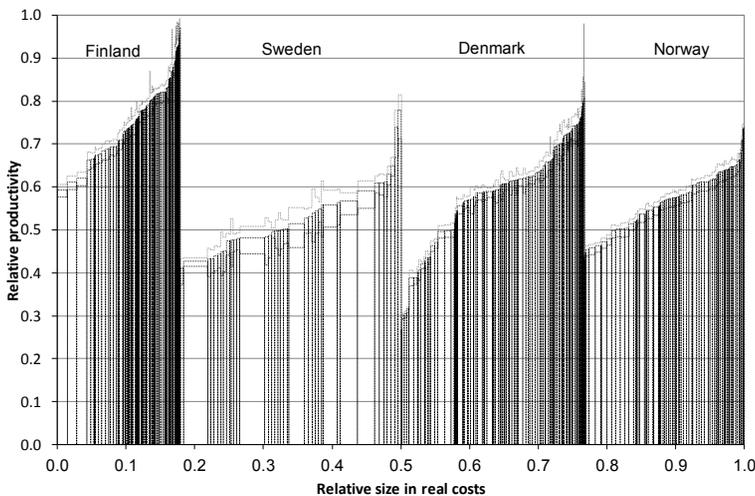
Source: Kittelsen et al. (2008).

The figure clearly shows how the efficiency development for the three last years almost coincides. The efficiency development of Norway showed a slight productivity increase from the years before the reform of 2002, to the years after the reform. A special analysis estimating the effect of the reform by using the other countries as controls showed that the

hospital reform in Norway had improved the level of productivity in a magnitude of approximately 4 percent or more (Kittelsen et al., 2008).

In updated data sets from the period 2005 to 2007, hospital efficiency in the Nordic countries was assessed using the same methods (Kittelsen et al., 2009). The dataset differs from the previous study by differences in aggregation where data from Sweden, and to some extent Norway, are based on regional authorities running several hospitals. Still, the country-level average differences in efficiency were surprisingly stable in time. Finland shows an 18 percent higher efficiency than Norway. The differences between Norway and the other two countries are not significant. The findings are presented in a Salter diagram (Figure 2).

Figure 2. Salter-diagram for efficiency-scores, somatic hospitals (95 % confidence-interval)



Source: Kittelsen et al. (2009).

As can be seen from the figure, the large units in each country dominate the segment with a low efficiency. We can also see that Sweden includes larger units, which is due to problems with access to disaggregated data. Hence, some of the larger units are county councils, not individual hospitals. The high efficiency scores in Finland are to a large extent found among the small local hospitals. University hospitals which are large in terms of volume and turnover generally show lower efficiency scores.

Further statistical analysis of the comparison between 2005 and 2007 was done in Kalseth et al. (2011), with decomposition of the productivity differences into cost efficiency, scale efficiency and country-specific effects. A positive association between efficiency and outpatient share was found. A long length of stay had a negative correlation with efficiency. The analysis showed large differences in the country-specific effects in terms of different frontiers. The overall frontier was determined by the Finnish hospitals. The analysis showed that the high productivity level in Finland is mainly due to domestic structural, financial and organizational factors that are common for all hospitals in each country. Hence, these country effects are essentially not caused by factors that each individual hospital could influence by itself to become more efficient. The technology frontier and possibility sets are determined at the country level. The analysis also showed small differences in scale and cost efficiency between countries. A conclusion from the decomposition analysis was that further research should focus on identifying the country-specific effects in terms of analysis of differences in financing, reimbursement and incentives, ownership structure, regulatory frameworks etc (Kalseth et al., 2011).

The relatively low efficiency for university hospitals has been shown in previous research. The provision of research and education by university hospitals interferes with patient care routines and inflates the costs of health care services, turning university hospitals into outliers in comparative productivity and efficiency analyses. In addition, these hospitals run most of the tertiary health services admitting and treating more severe cases than other hospitals. Their workload and special case-mix cannot be captured by the measurement of DRGs. A special study by the NHCSG was performed for all university hospitals in the Nordic countries (Medin et al., 2011). The results demonstrate significant differences in university hospital cost efficiency when the variables for teaching and research were entered into the analysis. Two major models were specified: the patient care production models (PC), including operating costs and patient care outputs, and the teaching and research (ToR) models, including costs for teaching and research as well as teaching and research outputs. The location of the frontier depends on whether the production function is assumed to exhibit constant returns to scale (CRS) or allows for variable

returns to scale (VRS). In Table 2, the result from the VRS-model, with the least restrictive assumption of returns to scale, is presented.

Table 2. Mean bootstrapped bias-corrected efficiency scores (standard deviations in brackets) and their confidence intervals, number of efficiency units per country (variable returns to scale)

Country	Patient care model			Teaching and research model		
	Mean eff	CI	Eff units	Mean eff	CI	Eff units
Denmark	0.90 (0.06)	0.82-0.95	6	0.95 (0.03)	0.90-0.98	7
Finland	0.95 (0.01)	0.91-0.98	5	0.97 (0.01)	0.93-0.99	9
Norway	0.88 (0.08)	0.84-0.90	1	0.93 (0.04)	0.90-0.95	8
Sweden	0.84 (0.10)	0.80-0.86	3	0.96 (0.03)	0.92-0.98	14

Source: Medin et al. (2011).

The number of efficient units per country is also presented in the table. Finland presents the highest average cost efficiency score in the patient care models and the ToR model, regardless of the production technology assumption, whereas the lowest cost efficiency estimates in the PC model are found among the Swedish university hospital observations, yielding the lowest country average. The Danish country averages are the second highest in the ToR model, and the Norwegian country averages are the second lowest. The production technology assumption has the largest effect on the Swedish country averages. The inclusion of costs and outputs for research and teaching yields some different results with a higher comparative efficiency score for Swedish hospitals in the ToR model than in the patient production models. Overall, the differences across countries are diminished in the ToR model.

Second-stage analyses were performed for both studies that were referred to. The efficiency scores in the study comparing all acute hospitals were regressed on a set of explanatory variables. One research question was to separate the effect of the Norwegian hospital reform from the effects of other structural, financial and organizational variables. The variables tested were case-mix index, changes of activity based funding (ABF) and length of stay (LOS) variations. A fixed hospital effect model was used, as random effects and OLS specifications are rejected (Kittelsen et al., 2008). The analysis showed that changes in ABF had no effect, that a longer LOS than expected (within each DRG) was associated with a lower efficiency and that case-mix did not have any significant effect.

The results of the second-stage analysis for the university study concerning the PC model confirmed that all country dummies, which reflect institutional and geographical differences that were not captured by other variables, have a statistically significant positive effect on the cost efficiency scores using Sweden as the reference country. The average operating cost per university hospital observation in Sweden was twice the size of the input averages in the rest of the Nordic countries, whereas the volume of patient care production is similar to that in Finland. The second-stage analysis also showed that the case-mix variables of importance in the PC models are the case-mix index (CMI) variable and a variable for super-specialized service. That the CMI variable had a negative effect on efficiency reflects that the DRG case-mix adjustment does not fully capture the variation in the material and apparently a higher CMI is more resource demanding. The second-stage DEA also shows that a higher ratio of doctor visits (outpatient) to inpatient discharges has a positive predictive effect on the cost efficiency scores, which is also expected since outpatient cases are less resource demanding compared to inpatient cases. The Swedish hospitals did not differ significantly from the other Nordic university hospitals in the ToR model. The Norwegian and the Danish university hospital observations present the highest means in the qualitative research indicators. Meanwhile, the total number of citations of Finnish and Swedish articles was higher (Medin et al., 2011).

3. Explaining the differences – some theoretical reflections

The presented studies did not depart from a theoretical standpoint or tried to test any specific hypothesis.⁶ Even if the Nordic systems share many similarities, there are different characteristics that could explain the observed productivity differences. This section discusses the findings in light of both economic theories of relevance, but also from a health policy perspective. In health economics and adjacent research areas, there is no consensus when it comes to establishing a theoretical framework for the structure of health care systems. Still, there are some essential characteristics of the health care market emphasized by health economists. Most of

⁶ An exception was the effort to analyze the effect of the Norwegian reform where a hypothesis of improved efficiency was supported (Kittelsen et al., 2008)

them are fundamental and not very controversial, but they are important to keep in mind as well as how they depart from classical economic theory.

In his classical article, Arrow (1963) emphasizes the state of uncertainty as a significant element of the health care market. Since illness is unpredictable, the demand or need for health service is uncertain for the individual. Just as uncertainty is an inherent characteristic, different insurance arrangements are derivations or responses to the market features (Evans, 1984). In modern societies, health insurance is provided by both private and public insurers, both establishing a third-party payer relationship with both providers and patients (consumers). Financial transactions between providers and patients are almost entirely replaced by a third-party payer arrangement. Other characteristics of the health market are the existence of information asymmetry, externalities and values about a fair distribution of services (equity). The information asymmetry problem exists between providers and insurers as well as between providers and patients.

The health systems in Northern and Western Europe today cover almost the entire population and the coverage includes most services. There is also a very small amount of money transactions between patients (consumer) and doctors/hospital (producer). Individual consumers are then shielded from financial consequences at the point of consumption. On average 70-75 percent of health spending pass through different types of third-party payers (insurers). The main difference we could observe is how the relationship between third party payers and providers is structured. Here, we find differences between tax-based systems and social health insurance systems. Tax-based systems have a tradition of vertical integration between financing and provision where most services have been produced in-house. In countries with social health insurance systems, contracting with independent private providers dominates the production side.

There are also some differences of importance within the tax-based systems like the Nordic countries. The Nordic countries belong to the tax-based group although they have different arrangements for the relationship among third-party payers, providers and patients. As presented above, the level of the financing system differs between Denmark and

Norway with a more centralized approach and Finland and Sweden with more of decentralization.

Besides taking the characteristics of the health market into account, there are other different theoretical contributions in economics that could give additional insights into the differences within public health systems such as in the Nordic countries.

Given a health system with a tax-based financing and an overall risk-sharing – the issue of how to organize the production side remains to be analyzed. The Nordic system has a tradition built on a vertically integrated system where the regional/local funders also provide most services. During the expansion of the systems in the 1960's and 1970's, many independent private facilities were socialized, such as nursing homes and pharmacies in Sweden. Many private providers became dependent on the tax contributions as the private payments were regulated or abolished. In a parallel process, most investment in new facilities took place in the public sector. During the 1980's, the efficiency of the public sector was questioned in many countries which gave rise to what was called the new public management or quasi-markets (LeGrand and Bartlett, 1993).

The quasi-market reforms took place within the public sector aiming at introducing market principles and competition within the public sector. The objectives were to promote higher efficiency and introduce consumer sovereignty within the sector. At the same time, the objectives of solidarity and equity remained through a public financing of the system. The quasi-markets have no exact definitions and there is a lack of theoretical framework for its function. Still, the literature on quasi-markets has to a large extent used economic theories of market failure problems and institutional economics.

Both the former structure of tax-based health care systems and the reforms from the 1980's and forward could be discussed in the light of different theoretical contributions. When it comes to the organization of the provision side, two major theoretical contributions are of relevance. First, institutional economics deals with the issue of the most efficient way of organizing production. The work by Williamson (1985) analyzes why some transactions take place in-house and others are contracted. Given the characteristics of the transactions, different forms of hierarchies and markets are considered as being appropriate. The choice between vertical integration in terms of in-house production could be ana-

lyzed from a transaction cost perspective. A crucial element is also how contracts are designed and written between purchasers and providers.

Another valuable contribution is the theories of non-market organizations. With the public choice school, the objectives of decision makers within the public sector are considered. The monitoring problems in the public sector could be analyzed from a principal-agent relationship. The public choice theory contributes with models on different arrangements and how decision makers are inclined to pursue their own sub-goals. The existence of discretionary behavior in the public sector might interfere with or have consequences for efficiency. As the in-house production mode dominates, but also differences within the health systems in the Nordic countries, both institutional economics and the public choice school could give insights into how to explain differences in hospital efficiency.

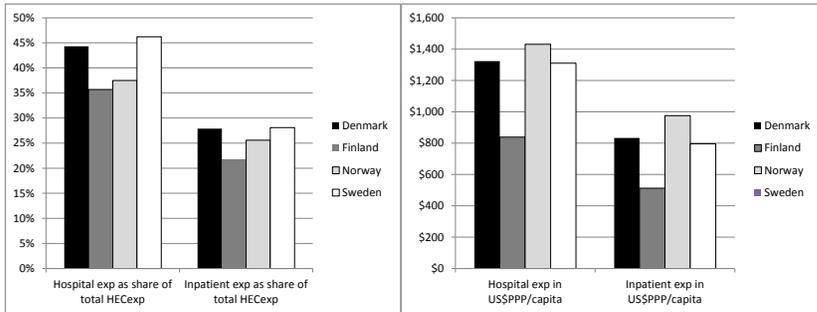
In addition to using theoretical contributions, this section also discusses the potential effect of the resource allocation system and the inherent incentives in each system. Furthermore, the issue of incorporating quality indicators in the efficiency analyses is discussed.

3.1 Health care spending and organizational structure

Spending pattern

The spending of resources in the health care sector shows some differences that could be of interest for analyzing the findings of hospital efficiency. Previous international studies have showed that countries allocate health care resources differently between hospital service, specialist outpatient service, primary health service, pharmaceuticals etc. Even if there is no consensus of an optimal mix of service, the size of the allocation to a specific sector could matter. In the case of hospital efficiency, the share and size of resources spent on hospital services could be crucial. These figures are not always comparable and should be interpreted with caution. Figure 3 shows the relative and absolute spending on hospital care in the Nordic countries based on both OECD figures and figures from the Nordic group (NHCSG).

Figure 3. Resources spent in hospital service and inpatient treatment, Nordic countries, year 2004



Source: OECD Health Database.

As shown from the figure, there is a pattern that Finland, irrespective of source, devotes a smaller share to the hospital sector than the other three Nordic countries. The Swedish system has been structured as a hospital-oriented system for a long time. The higher utilization of hospital service in Sweden could be justified by a somewhat older population than in the other countries. The data from the NHCSG show a higher cost per capita in Sweden for most hospital services, but not a correspondingly higher volume of outputs. Hence, with a larger amount spent on hospitals service, a higher utilization (output) or performance is necessary for keeping efficiency at an equal level. As has been shown above, this is not the case and both the utilization and the efficiency scores are lower for Sweden, but also for Denmark and Norway as compared to Finland. The extra resources for the hospital sector do not yield a corresponding extra output.

Obviously, Finland devotes less resources to the health care sector in total than the other countries. The cost per capita is lower for both total hospital expenditures and inpatient care compared to the neighboring countries. The overall lower cost level, in particular the lower hospital costs, is partly an effect of the economic crisis in the 1990's. Employment in the Finnish public sector was drastically affected during the economic crisis with reductions of staff. The employment in local governments was reduced from 434 000 employees in 1990 to 394 000 in 1995. By the year 2000, the numbers were still lower than before the crisis (410 000).

Table 3. Physician intensity and distribution across specialties, Nordic countries

	GPs/1000 inh.	GPs/1000 inh.	GPs a % of all physi- cians	Number of physicians in somatic specialties /1000 inh.		GP-visits /inh.	Specialist- visits/inh.
				Medical specialties	Surgical specialties		
	OECD	NOMESKO	OECD	OECD	OECD	NOMESKO	NOMESKO
Denmark	0.69	0.71	22%	0.52	0.49	3.65	0.78
Finland	1.01	0.71	34%	0.58	0.37	1.28	1.28
Norway	--	0.91	24% appr.	0.52	0.43	--	--
Sweden	0.59	0.53	17%	0.75	0.56	1.18	1.37

Source: OECD Health Database and NOMESKO.

Another indicator for comparison on the input side is the availability of staff in hospital service, but also the distribution of staff between sub-sectors. In total, the number of doctors per capita is somewhat higher in Norway and Sweden (3.4/1 000 inh.) than in Denmark (3.2/1 000 inh.) and Finland (3.0/1 000 inh.). As shown in Table 3, the figures for the doctor per capita ratio and the share of doctors differ between sub-sectors.

As seen in Table 3, the General practitioner (GP)/population ratio is lowest in Sweden and highest in Finland. Sweden devotes a smaller share of its doctors to the GP sector and Finland the highest. Sweden deviates from the rest of the Nordic countries by having more visits in specialist service than in primary health care. The Finnish consultations to primary health service also include a sector of occupational doctors. The overall picture shows that Sweden devotes a larger share of its health expenditures and doctors to the hospital sector, without a correspondingly higher rate of performance or utilization. To sum up, we can notice certain observations as a generally lower level on health care spending in Finland, but also a more hospital-orientation of resources and allocation of doctors in Sweden.

Organization of purchasers and providers

The decentralized Nordic health systems have been dominated by relatively autonomous regions and municipalities with their own political decision bodies. For a long period of time, the financing and provision functions were integrated under a sole ownership and resources allocated through internal budgets. Unlike the other Nordic countries, Finland has not yet implemented any reforms to establish larger municipalities and regional levels.⁷

The new public management or the quasi-market systems were first implemented in the UK in the early 1990's. Sweden was one of the first countries after the NHS in the UK that adopted the purchaser-provider split model within some public services. Several county councils split their organization into a purchasing side aiming at focusing on the consumer (patient) interest and a provider side focusing on the running of health facilities such as hospitals and primary health centers. The re-

⁷ In Finland during the 1990's, many recommendations were made in order to increase the population base for units responsible for health services (Häkkinen and Jonsson, 2009). The skeleton law passed by the parliament in 2007 and more recent proposals aim at increasing the population base for the municipalities and centralizing certain services.

centralization reforms in Denmark and Norway have also incorporated some ideas based on a purchaser-provider split.

In Finland, there were no special purchasing units set within the municipalities and the public hospitals were not given an autonomous status, but are still directly managed by the owners (the hospital districts). Still, the traditional *at arm's length* relationship between the municipalities and the hospitals could be seen as a transaction that shares similarities with a so-called *quasi-market* and a purchaser-provider split. Even if the municipalities have not exercised their role as active purchasers, they have been faced with difficult resource allocation decisions. As purchasers of hospital service and providers of primary health centers, a trade-off has to be made between the different types of services. At the same time, they could indirectly use their role as part-owner to monitor the hospital. Yet, the heterogeneity of the pricing system makes it difficult to compare hospital service and its prices.

The purchaser-provider split (or the quasi-market structure) could be arranged in different ways. The Finnish model differs from the other countries through a transaction between public authorities. The other Nordic countries have implemented the transactions internally with purchasers without own production (no fund-holding) and autonomous providers. This type of quasi-markets does often create a bilateral monopoly situation, especially in the hospital sector. In a bilateral monopoly, prices and outputs will be determined by forces like the bargaining power of both purchasers and providers. The outcome would then be determined by the power balance between the actors. As hospitals often have the central role among health providers, they might have a stronger bargaining power than the purchasers. If we apply the theories of imperfect markets, we should look at the monopoly status of the Finnish hospitals with a market of several buyers. There is little encouraged competition between hospitals and patients do not have the right to freedom of choice as in the other Nordic countries. Hospitals would then have the possibility of charging monopoly prices for their services, but not of influencing the volume of patients. The municipalities could use their power as owner to control supply and also to put cost containment measures on hospitals.

The small size of the municipalities as payers makes them vulnerable to high variations in hospital costs. Since the hospital service is used by few individuals at high variable costs, the municipalities with a small

population base are faced with a large uncertainty concerning hospital costs. A study by Mikkola et al. (2003) analyzed the financial risk and found that the size of the municipality was the main determinant. One way of coping with the risk is to keep control over the utilization of hospital services.⁸ The purchasers in the other Nordic countries are organized as larger units and could spread this financial risk over a larger population. The incentive for reducing the use of hospital service might then be lower. As the municipalities do not only fund the hospital service but also run facilities such as primary health service and long-term care, they do not only have an incentive for a short hospital stay, but also for treating patients in their own institutions.

In order to identify mechanisms within the public health care systems in the Nordic countries that could contribute to explain efficiency differences, models from the non-market literature could be helpful. The monitoring problem in the public sector has its analogue in private business. The existing theories of the managerial firm have influenced the public choice analysis of government bureaucracy. The common underlying assumption is that all parties act as utility-maximizers, thereby implying the existence of different forms of discretionary behavior in all organized activities. Managerial discretion will exist in both private and public firms. Politicians are assumed to act as vote-maximizers. Managers and professionals in all organizations are inclined to pursue their sub-goals. Mueller (1991) states that the discretionary behavior problem is similar in profit-oriented and non-profit oriented bureaucracies. The budget-maximization model launched by Niskanen (1971) follows a sales-maximization model from the for-profit sector, stating that compensation schemes are often based on the size of turnover and the number of employees. Borcharding (1977) noticed a tendency for governments to produce service in-house. According to that study, this effect is more pronounced among local governments. De Alessi (1969) claims that public managers exhibit preferences for labor and capital. Orzechowski (1977) concluded that public agencies experienced productivity losses, operated at costs above those of private firms, and showed a significantly greater labor-to-capital ratio.

⁸ Each hospital district in Finland has a risk equalisation system for very resource intensive patients, which means that when the patient costs per year exceed a limit, the system compensates the costs.

All Nordic health systems are built on a tradition of integrated public health services. From a macro perspective, the systems seem to have a similar integrated organization where non-market theories give limited help to distinguish the differences among the Nordic countries. Still, a closer look at details could be a link to theoretical contributions. The tendency to produce service in-house at local governments could give some explanations for the observed differences. In all Nordic countries, apart from Finland, the purchasing units do not run their own services. Hence, in Denmark, Norway and Sweden, the task of the purchasers is purely to contract with public and private providers. The Finnish municipalities, which are local governments, also have the choice to spend resources on local in-house services or contract with hospital and specialist services. If local managers have preferences for in-house services, a larger share of health spending will take place at the local level and priority will be given to primary health services.

There might be different reasons why local governments are prone to organize production in-house. Here, the proposed behavior from the public choice school could contribute to this. The vote-maximization paradigm could imply that local governments encourage high employment at local facilities in order to get re-elected. The budget-maximization incentives will, at the same time, reinforce the preferences for in-house production at local levels. The evidence from the GP-fund-holding and the primary care trusts (PCTs) in the NHS in England shows that fund-holders restrict the use of external or contracted services. The utilization of prescribed drugs was decreased with GP-fund-holders as compared to non-fund-holders (Propper, 2012). The GP fund-holders, which also run their own production, had to balance their own GP service against hospital service. The Finnish system shares some characteristics of a fundholding system as the municipalities have to pay (at arm's length) for hospital services. If municipalities also try to maximize their budget and have a preference for in-house production,⁹ this might lead to a short length of stay and a pressure to lower the hospital costs. As they run their own production for GP service and long-term care, they both have an incentive to spend more on their own in-house production and act for an early dis-

⁹ The DRG payment scheme was found to be a better protection against financial risk than a reimbursement based on bed-days (Mikkola et al., 2003). The DRG-system has become more common which will decrease the financial risk of the municipalities.

charge of patients from hospitals. As they have their own capacity for health service in terms of running health centers on an in-house basis, such transfers are facilitated.

Local government expenditure in Finland accounts for over 30 percent of total public sector expenditure and 2/3 of public consumption. The municipalities have several main responsibilities in the public sector besides health care. The other main areas are social welfare and education and culture. The relative size of these sectors is approximately 30 percent each. The local authorities have considerable freedom to decide how to spend their money. These decisions also force them to make priorities across the different public sub-sectors. During the economic recession in the 1990's, the municipalities got an increased responsibility for health care, which forced them to focus on cost containment as the main starting point for their actions. It is widely considered that this experience of saving and focusing on cost containment had a considerable effect on the behavior of the municipalities during the late 1990's and even the early 2000. This is different from the Swedish county councils, also with a decentralized delegation for health service, where the health service amounts to 80 percent of their expenditures. Hence, the Swedish county councils do not have to balance health care spending against other public service commitments.

There are both theoretical and empirical contributions that somewhat support the behavior of local government agencies such as containing costs for contracted services and perhaps promoting in-house provision at the local level. The municipalities in Finland act as fund-holders and can put pressure on the hospitals to improve efficiency and minimize the hospital costs. The incentives to control costs for hospital service and certify an efficient production might also be explained from a Tiebout competition perspective.¹⁰ The municipalities in Finland are rather small and moving across the borders is more likely and possible as compared to the larger regions in the other Nordic countries. Hence, the Finnish municipalities are closest to a Tiebout competition arrangement among the

¹⁰ The Tiebout model departs from the provision of public goods and not from private goods such as health service. The model states that the choice process of individuals and residents will determine the provision of local public goods in accordance with the preferences of residents. The model rests on assumptions that residents can move from community to community at no cost, that there is a large number of communities to choose from, complete information and sets of possible choices, etc.

Nordic countries. The size of the regions is rather large in the other countries involving high transaction costs for moving. Moreover, in Finland there are limitations and it is mostly the active working age population who has the best possibilities of moving across borders. Yet, as the municipalities are in charge of such areas as health service, social service and education, the allocation of resources will affect all sectors. Also from a public choice perspective, the discretion for in-house production could explain this behavior. The municipalities want to take care of most patients and spend most of their health care resources in their own facilities (PHC and long-term care). Even if the competition among municipalities does not follow the original Tiebout competition paradigm, the Finnish local government structure is closest to the conditions for a Tiebout competition (many municipalities, etc.) among the Nordic countries.

3.2 Resource allocation and incentives

The principle for resource allocation in the public sector is a puzzle and differs from market transactions. The traditional model has been based on budgets where the historical cost has often set an economic framework. The disjunction between costs and revenues (and outputs) gives a scope for misallocation of resources. By removing this link, the public providers do not operate under pressure from the environment to choose technologies which minimize the costs.

Still, the public sector shares some of the characteristics of hierarchies within private firms.¹¹ Williamson (1997) describes from a transaction cost perspective the alternative modes of governance structure and argues that hierarchies use low-powered incentives, administrative control, and resolve disputes within the firm. He also refers to an intermediate mode of governance called 'hybrid'. With the hybrid mode, the incentives are combined with administrative control.

Most public health care systems and to some extent also social health insurance systems are characterized by compulsory taxes, limited consumer choice and lack of competition. In the tax-based system, most provision takes place in-house in government institutions. In the last twenty

¹¹ A major difference is the competitive environment that private firms are faced with. A profit-maximization firm in a competitive market will choose the most efficient mode of governance structure.

years, several governments implemented market-oriented reforms such as competitive tendering and increased patient's freedom of choice by different 'voucher'-systems. The creation of internal markets¹² in the British NHS in the early 1990's was later introduced in some county councils in Sweden. The aim was to achieve a more efficient production of service without endangering the solidarity principle of financing health care. In the UK and the Nordic countries, the reform focused on the delivery side (Propper, 2012; Rehnberg, 1995). Purchasing units were set up in order to promote consumers' interest and providers were given a more autonomous status. The payment system was changed from a traditional budget-based resource allocation to various activity based fundings of providers.

Following the US experience in the Medicare program, several European countries introduced DRGs as a case-based payment method. Among the Nordic countries, this was first done in some of the Swedish county councils. A comparison from the mid-1990's showed that those county councils with a purchaser-provider split had a higher efficiency level compared with those relying on a budget-system (Gerdtham et al., 1999). Norway has used DRGs in activity based financing since 1997 and Denmark introduced DRGs as a marginal payment in 1999, but has increased the use to cover 50 percent of income in 2007. In Finland, the DRG-system is not used for resource allocation but mainly as a method of collecting payments from municipalities, i.e. as a billing instrument that is equitable between municipalities (Kautiainen et al., 2011). As previously mentioned, the transactions between municipalities and hospital districts in Finland could be characterized as a purchaser-provider split model, although a formal activity-based funding system is not in place.

In one of the analyses from the Nordic collaboration project, an explanatory model was tested where activity-based funding was one of the potential determinants for efficiency (measured by DEA). The initial purpose of the analysis was to test the effect of the Norwegian ownership reform using the other Nordic countries as controls. The DEA regression result shows that productivity had increased, but no significant effect of changes in activity based financing was found (Kittelsen et al., 2008). Previous studies of variations within countries in both Norway and Sweden show support for a higher efficiency among regions/counties with

¹² Sometimes the concept of 'Quasi-markets' is used which could be seen as a synonym to the hybrid mode in Williamson (1999), but applied to the public sector.

activity-based funding (Biorn et al., 2003; Gerdtham et al., 1999). Hence, the higher hospital efficiency in Finland could hardly be explained by incentives due to activity-based funding. On the contrary, these instruments have actually been used for a longer period in Denmark, Norway and Sweden.

A more important aspect of the difference in resource allocation between countries is most likely the relationship and transactions between the government levels. The transactions of resources in Finland are organized through a more distinct separation of purchasers and providers, whereas Denmark, Norway and Sweden have implemented this split internally to various extents. Following the reforms at the beginning of the new millennium, Denmark and Norway have centralized financing and provision of the hospital sector, whereas Finland and Sweden keep their decentralized structure. Still, in Finland, the transactions take place between legal entities, the municipalities and the hospital districts, whereas the Swedish purchaser-provider split takes place within one legal entity, the county council. Finally, we could observe that Finland has not implemented the autonomous status of the providers, nor has it used activity based funding to the same extent as its neighboring countries. Hence, the difference in hospital efficiency is not likely to be explained by differences in competition degree or the incentive structure arising from payment mechanisms. Table 4 shows the number of purchasers and providers in the four countries and compares the market structure with the efficiency scores.

Table 4. Market structure in the Nordic countries

Country	No. of funders	No. of providers	No. of hospitals	Funder/ Provider ratio	Market structure	Fund-holding	Average DEA-score*
Denmark	5	5	54	1.0	"Bilateral monopoly"	No	0.66
Finland	415	20	38	20.8	Monopoly	Yes	0.80
Norway	4	4	43	1.0	"Bilateral monopoly"	No	0.57
Sweden	21	21	63	1.0	"Bilateral monopoly"	No	0.49

Source: Own construction.

Note: *) Year 2004 for Finland, Norway and Sweden, year 2002 for Denmark.

Both Finland and Sweden show the most decentralized structure for both financing and provision. Denmark and Norway entered a re-

centralization development at the beginning of the millennium. From a fiscal perspective, the central governments in Finland and Sweden cannot exercise a direct cost control, but an effective cost control on local and regional expenditures is achieved based on annual decisions of state subsidies. One observation from Table 4 is that the regional division for purchasers and providers coincides in Denmark, Norway and Sweden. This creates a bilateral monopoly situation within the health authorities. The contractual agreements are also not legally binding in these transactions. Overall, little analysis has been made about the bargaining situations in these internal relationships. The presence of soft-budget constraints is one issue that could occur in such internal arrangements (Kornai et al., 2003). Finland differs in two aspects from the other countries. First, there is a contractual relationship between two administrative entities, the municipalities and the hospital district. They are legally independent partners, although the municipalities also have the position as part-owners of the hospital districts. A potentially more important aspect and difference between Finland and the other countries is the fundholding which has already been discussed. Further analysis should focus on how the trade-off decisions between the purchase of hospital services and the allocation of resources to in-house primary care and long-term care are made among the Finnish municipalities.

Competition from the private sector is very limited in all countries. The only for-profit acute hospital is located in Stockholm, Sweden and, in addition, there are small numbers of private hospitals providing elective treatment in all countries. Hence, the hospital market is almost entirely owned and regulated by the public authorities; a competitive environment as in the primary health sector in Norway does not exist.

3.3 Quality differences

The literature of hospital behavior could be divided into different classes. The most relevant models deal with the behavior of non-profit firms, which could theoretically also be applicable to public hospitals. Most analysts posit an objective to the hospital decision makers. A classical model was proposed by Newhouse (1970) where the hospital's objective was to maximize the utility of the decision makers. The hospital is described as a very complex organization with elements and interactions

between decision makers. Newhouse's theory departs from the nonprofit hospital behavior in the US context, where the main decision makers are the trustees, the management or the administration, and the physician staff. Different decision-makers are assumed to value special objectives that could be considered as quantitative or qualitative. In a public health care setting such as the Nordic hospitals, the main decision makers would be politicians ("trustees"), managers and physicians. Both politicians and managers are assumed to value quantity outputs as the numbers of patients treated, short waiting lists etc. They could also be assumed to value quantity as the number of admissions and patient cases as it justifies the staff structure and ultimately the budget and revenues. Physicians are assumed to value access to modern advanced technologies, survival rates, conducting research outputs etc. which are considered as quality objectives. Of course, all decision makers could have different preferences along the quantity-quality trade-off. The hospitals will then select a combination of quantity and quality that maximizes utility. They also face a budget constraint, as they have to pay for expenses and cannot run a deficit.¹³

The hospital does not simply maximize quantity or maximize quality, but instead selects a combination of quantity and quality that maximizes utility. Newhouse combines the quantity and quality maximization objectives in one model with an equilibrium for every quality level. This trade-off could differ depending on the internal power struggle between decision-makers depending on how they value different quantity or quality outputs. Given the scope for discretionary behavior, this trade-off might differ across countries.

¹³ This could be questioned in the public sector, see the theories concerning soft-budget constraints (Kornai et al., 2003).

Table 5. Selection of quality indicators linked to hospital performance in the Nordic countries

Country	Revision burden for total hip replacements	Five-year relative cancer survival adjusted for age and case-mix (w standard errors)		Mortality within 28 days* of myocardial infarction. Hospital admission (and total), percentage		Avoidable deaths per 100 000 population, 2001
		Men	Women	Men	Women	
Denmark	15% (1995-1999)	36.7 (0.5)	53.5 (0.5)	30 (56)	44 (65)	124
Finland	24% (1995-1999)	46.2 (0.4)	56.9 (0.3)	21 (48)	24 (39)	84
Norway	15% (1987-1999)	43.2 (0.4)	55.8 (0.3)	--	--	72
Sweden	11% (1992-2000)	46.4 (0.3)	57.9 (0.3)	10-11 (32-44)	14-17 (32-46)	68

Source: SHPR, OECD Health Database, NOMESKO.

Note: *) Mortality within 28 days of myocardial infarction at 31 centers in WHO's MONICA project. Total percentage of deaths within 28 days and percentage of deaths within 28 days of hospital admission (alive at time of admission).

The Newhouse model could be used for interpreting the differences in hospital efficiency between the Nordic countries. The study by Medin et al. (2011) that was previously referred to indicates that the Swedish university hospitals perform better when research and teaching outputs, together with the costs thereof, are added into the efficiency analysis. Even if the study only analyzes the efficiency of university hospitals, the findings could be interpreted as the hospitals' quantity/quality trade-off differing between the countries. One possible explanation for the differences is a greater focus on quality and research-oriented performance (and less on quantity performance) in Sweden. The Norwegian figures also become more positive when research and teaching variables are included in the DEA-model.

Even if research should not be seen as a perfect proxy for quality, future research on efficiency gives possibilities to incorporate quality indicators. The comparative research concerning the quality difference in health care is very limited. Still, for some medical procedures, there are published comparisons. In addition, some international organizations and agencies have compared the survival rates of different procedures. Some performance indicators are presented in Table 5.

The table includes quality indicators that to a large extent are controlled and a function of hospital performance. Other measures such as life expectancy and infant mortality are influenced by several external factors outside the control of hospitals. As we can observe, there are differences across the Nordic countries both for technical indicators such as hip revision, but also survival and mortality rates. A recent study on inpatient hospital mortality showed that Swedish hospitals are performing better than Finnish hospitals in treating acute myocardial patients, whereas the opposite prevailed in treating stroke patients (Häkkinen et al., 2012). Another relevant indicator is avoidable death which is aimed at showing what health services can achieve through prevention and direct interventions. Among the Nordic countries, Sweden reports the lowest rate of avoidable deaths. By incorporating quality indicators, the efficiency scores might be affected.

One experience of health care markets is the difficulties in establishing a competitive behavior between providers. Especially an efficient price-competition seems difficult to achieve. The funder and purchasers of health service in tax-based systems have a clear monopsony power that

is used through decisions of in-house production and contracting-out, public procurement and setting incentives for reimbursement systems. The performance of different strategies for organizing the provision of health services shows vague results. The objective of cost containment has been successful, but the evidence of efficiency improvements is unclear. More lately, several government and health authorities have encouraged and supported benchmarking systems as a measure for improving efficiency and quality. In the Nordic countries, this development has had a somewhat different focus. In Finland, a hospital benchmarking system was piloted in 1996 and two years later, all hospital districts in Finland participated in the project (Linna and Häkkinen, 2008). Since the year 2006, this information has also been public.¹⁴

In Sweden, different initiatives have been taken by both professional medical groups and national authorities such as SALAR and National Board of Health and Welfare. The development started on a voluntary base by medical professional groups where the earliest started in the 1980's. In order to stimulate comparisons and contribute to a greater openness concerning results and costs, the central government started to give grants for the so-called quality-register. Today, a yearly report of about 100 performance indicators is published with measures of quality within four areas: medical results, patient experiences, availability and costs. The Swedish benchmarking system differs from the Finnish system in several respects. Most of the indicators are results at the regional level (county council). Only 10-20 indicators are at the hospital level. More important concerning the hospital efficiency is that the Swedish benchmarking system does not provide any cost-efficiency measures at the hospital level, since cost figures at the hospital level are lacking. Hence, whereas the Finnish benchmarking system during the periods for the productivity analysis has focused on hospital cost-efficiency, the Swedish system has focused on quality indicators. More benchmarking and disclosure of information about provider performance could be one way of promoting efficiency. The experience of benchmarking in both Finland and Sweden has led to public discussion and there are signs of improvement and a reduction of variations in performance across providers.

¹⁴ More recently, after the study period referred to, a project focusing on quality and efficiency has been launched in Finland (Häkkinen, 2011).

4. Concluding remarks

Comparative studies of the performance of health systems is a source for identifying and explaining differences in costs, outcome and efficiency. Acute short-term hospitals are the major resource user in the health care sector and have a significant role for advanced treatment. In this paper, the findings from the Nordic collaboration on productivity differences across acute hospitals have been presented and discussed. As the four countries share many administrative tools and use common standards for data collection, unique cross-country comparisons are possible. The results suggest that there was a markedly higher average hospital efficiency in Finland as compared to Denmark, Norway and Sweden. The efficiency has been analyzed for two time periods with somewhat different datasets. In addition, a special study of the university hospitals including research and teaching variables has been performed. The findings are rather robust in terms of the ranking of the countries.

A further analysis showed that country-specific effects are not correlated with the explanatory variables tested. This means that these country effects cannot be changed by individual hospitals but must be linked to the structure of financing, regulatory framework, organizational arrangements etc. in each country. The explanations of findings are discussed along different theories and possible reasons for the observed differences. Although no clear explanations are argued for, a number of hypotheses for further research are identified. The markedly higher efficiency levels among the Finnish hospitals do not seem to be explained by differences in the use of market mechanisms and reimbursement systems. The Finnish system has not implemented any performance-based payments or internal market mechanisms. The method and arrangements for the allocation of resources in Finland between different health services and also the trade-off against other public duties are proposed as major differences in relation to the neighboring countries. The combined role as purchaser and provider at the municipality level is also proposed as an important aspect of resource allocation within the health sector. The paper argues for a closer analysis of the impact of fund-holding, contractual relations and incentives between public governments as well as including quality indicators in the efficiency measure.

References

- Arah, O., Westert, G., Hurst, J. and Klazinga, N. (2006), A conceptual framework for the OECD Health Care Quality Indicators Project, *International Journal for Quality in Health Care* 18, 5-13.
- Arrow, K.J. (1963), Uncertainty and the welfare economics of medical care, *American Economic Review* 53, 941-973.
- Biorn, E., Hagen, T.P., Iversen, T. and Magnussen, J. (2003), The effect of activity-based financing on hospital efficiency: A panel data analysis of DEA efficiency scores 1992-2000, *Health Care Management Science* 6, 271-283.
- Borcherding, T.E. (1977), One hundred years of public spending, 1870-1970, in T.E. Borcherding (ed.), *Budgets and Bureaucrats: The Sources of Government Growth*, Duke University Press, Durham NC.
- Charnes, A., Cooper, W.W. and Rhodes, E. (1978), Measuring the efficiency of decision-making units, *European Journal of Operational Research* 2, 429-444.
- De Alessi, L. (1969), Implications of property rights for government investment choices, *American Economic Review* 59, 13-24.
- Evans, R.G. (1984), *Strained/Mercy: The Economics of Canadian Health Care*, Butterworths, Toronto.
- Gerdtham, U., Rehnberg, C. and Tambour, M. (1999), Estimating the effect of internal markets on performance in Swedish health care, *Applied Economics* 31, 935-945.
- Hansen, K.K. and Zwanziger, J. (1996), Marginal costs in general acute care hospitals: A comparison among California, New York and Canada, *Health Economics* 5, 195-216.
- Hurst, J. and Jee-Hughes, M. (2001), Performance measurement and performance management in OECD Health systems. Labor market and social policy, *Occasional Papers* 47, OECD.
- Häkkinen, U. (2011), The PERFECT project: Measuring performance of health care episodes, *Annals of Medicine* 43 (Suppl 1), S1-3.
- Häkkinen, U., Cots, F., Geissler, A., Kapiainen, S., Or, Z., Peltola, M., Rosenqvist, G., Rättö, H., Serdén, L. and Sund, R. (2012), Quality, cost, and their trade-off in treating AMI and stroke patients in European hospitals, manuscript, Centre for Health and Social Economics, National Institute for Health and Welfare, Helsinki.
- Häkkinen, U. and Jonsson, P.M. (2009), Harnessing diversity of provision, in J. Magnussen, K. Vranæk and R.B. Saltman (eds.), *Nordic Health Care Systems: Recent Reforms and Current Policy Challenges*, Open University Press, Berkshire.
- Häkkinen, U. and Joumard, I. (2007), Cross-country analysis of efficiency in OECD health care sectors: Options for research, *Economics Department Working Papers* 554, OECD, Paris.
- Iversen, T. (2011), *Vägval i vården – en ESO-rapport om skillnader och likheter i Norden*, Finansdepartementet, ESO 2011:7, Stockholm.
- Joumard, I., André, C. and Nicq, C. (2010), Health care systems: Efficiency and institutions, *OECD Economics Department Papers* 769, Paris.
- Kalseth, B., Anthum, K., Hope, Ø., Kittelsen, S. and Persson, B. (2011), *Spesialisthelsetjenesten i Norden, Sykehusstruktur, styringsstruktur og lokal arbeidsorganise-*

- ring som mulig forklaring på kostnadsforskjeller mellom landene, SINTEF, Trondheim.
- Kautiainen, K., Häkkinen, U. and Lauharanta, J. (2011), Finland: DRGs in a decentralized health care system, in R. Busse, A. Geisler and M. Quentin Wiley (eds.), *Diagnosis-Related Groups in Europe: Moving towards Transparency, Efficiency and Quality in Hospitals*, European Observatory on Health Systems and Policies Series, Brussels.
- Kittelsen, S., Magnussen, J., Anthun, K., Häkkinen, U., Linna, M., Medin, E., Olsen, K. and Rehnberg, C. (2008), Hospital productivity and the Norwegian ownership reform – A Nordic comparative study, Working Paper 2008:10, Health Economics Research Programme, University of Oslo.
- Kittelsen, S., Anthun, K., Kalseth, B., Kalseth, J., Halsteinli, V. and Magnussen, J. (2009), En komparativ analyse av spesialisthelsetjenesten i Finland, Sverige, Danmark og Norge: Aktivitet, ressursbruk og produktivitet 2005-2007, juli 2009, SINTEF Helsetjenesteforskning og Frischsenteret, Oslo.
- Kornai, J., Maskin, E. and Roland, G. (2003), Understanding the soft budget constraint, *Journal of Economic Literature* 42, 1095-1136.
- LeGrand, J. and Bartlett, W. (eds.) (1993), *Quasi-Markets and Social Policy*, MacMillan Press Ltd, Hong Kong.
- Linna, M. and Häkkinen, U. (2008), Benchmarking Finnish hospitals, *Advances in Health Economics and Health Services Research* 18, 179-190.
- Linna, M., Häkkinen, U., Peltola, M., Magnussen, J., Anthun, K.S., Kittelsen, S., Roed, A., Olsen, K., Medin, E. and Rehnberg, C. (2010), Measuring cost efficiency in the Nordic hospitals – A cross-sectional comparison of public hospitals in 2002, *Health Care Management Science* 13, 346-357.
- Magnussen, J., Vrangbaek, K. and Saltman, R. (2009), *Nordic Health Care Systems: Recent Reform and Current Policy Changes*, Open University Press/McGraw Hill, London.
- Medin, E., Anthun, K., Häkkinen, U., Kittelsen, S., Linna, M., Magnussen, J., Olsen, K., Peltola, M., Roed, A. and Rehnberg, C. (2011), Cost efficiency of university hospital teaching, research and health care, *European Journal of Health Economics*, 12, 509-519.
- Mikkola, H., Sund, R., Linna, M. and Häkkinen, U. (2003), Comparing the financial risk of bed-day and DRG based pricing types using parametric and simulation methods, *Health Care Management Science* 6, 67-74.
- Mobley, L. and Magnussen, J. (1998), An international comparison of hospital efficiency. Does institutional environment matter?, *Applied Economics* 30, 1089-1100.
- Mueller, D.C. (1991), *Public Choice II*, Cambridge University Press, Cambridge.
- Newhouse, J. (1970), Toward a theory of nonprofit institutions: An economic model of a hospital, *American Economic Review* 60, 64-74.
- Niskanen, W.A. (1971), *Bureaucracy and Representative Government*, Aldine-Atherton, Chicago.
- Orzechowski, W. (1977), Economic models of bureaucracy: Survey, extensions, and evidence, in T.E. Borcherding (ed.), *Budgets and Bureaucrats: The Sources of Government Growth*, Duke University Press, Durham NC.
- Pedersen, K.M. (2004), Health care reforms in Denmark and Norway, in B. Jönsson, G. Arvidsson, L. Levin and C. Rehnberg (eds.), *Hälsa, vård och tillväxt*, SNS, Stockholm.

- Propper, C. (2012), Competition, incentives and the English NHS, *Health Economics* 21, 33-40.
- Rattsø, J. (2002), Fiscal controls in Europe. A summary, in B. Dafflon (ed.), *Local Public Finance: Balanced Budget and Debt Control in European Countries*, Edward Elgar, Cheltenham.
- Rehnberg, C. (1995), The Swedish experience with internal markets, in M. Jérôme-Forget, J. White and J.M. Wiener (eds.), *Health Care Reform Through Internal Markets – Experience and Proposals*, Brookings/IRPP, Montreal.
- Tiebout, C.M. (1956), A pure theory of public expenditures, *Journal of Political Economy* 64, 416-424.
- Williamson, O.E. (1985), *The Economic Institutions of Capitalism*, The Free Press, New York.
- Williamson, O.E. (1997), Public and private bureaucracies: A transaction cost economics perspective, *Journal of Law, Economics and Organization* 15, 306-342.

Comment on Rehnberg and Häkkinen: Productivity differences in Nordic hospitals: Can we learn from Finland?

Thorvaldur Gylfason*

This paper constitutes a useful exercise in benchmarking, asking a simple, basic question: Can we learn from those who seem to produce better results than we do and, if so, what?

Consider this example. In an international examination of reading, mathematics, and science given every three years since 2000 to about 5 000 15-year-olds around the world, Finnish students have consistently achieved stellar results. The examinations are held under the auspices of the Program for International Student Assessment (PISA), a worldwide study by the Organization for Economic Co-operation and Development (OECD).

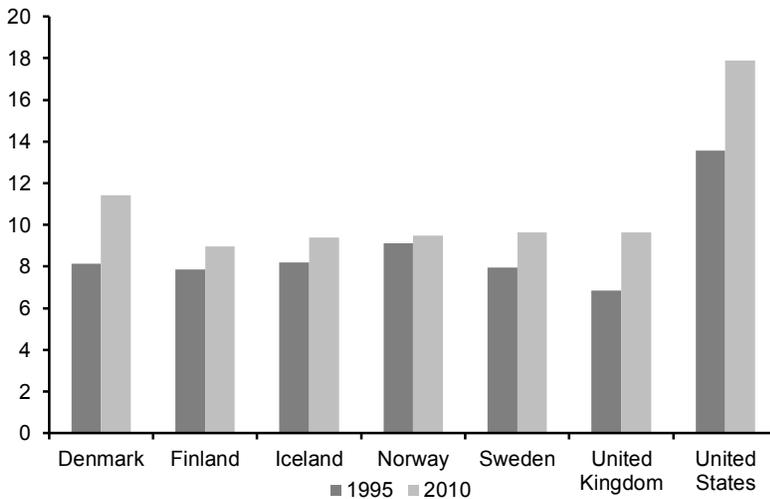
Why do the Finns do so well? There is no shortage of possible explanations. Some observers attribute their success to broader curricula, smaller classes, and better training, pay, and treatment of teachers than in other countries. Others are not so sure. The advantage of the PISA tests is that they are designed to measure the quality of education by output rather than by input such as school enrollment rates, years at school, and expenditure on education.

By the same token, Rehnberg and Häkkinen's paper is triggered by the rather striking empirical finding that average hospital efficiency in Finland appears to be markedly higher than in Denmark, Norway, and Sweden. The paper goes on to explore possible reasons for the observed dif-

* University of Iceland, gylfason@hi.is.

ferences, suggesting a number of conceivable explanations for further research. It is interesting and, to some, no doubt also surprising that the findings apparently cannot be traced to differences in the use of market mechanisms and reimbursement systems. It might seem tempting to suspect that a decentralized health care sector would produce efficiency gains that could help explain the superior efficiency of Finnish hospitals, but that particular explanation appears doubtful because Denmark and Norway have centralized their hospital sector, whereas Finland and Sweden have kept their decentralized structure and yet, as Rehnberg and Häkkinen show, Norway seems more efficient than Sweden. They argue for a closer analysis of the impact of fund-holding, contractual relations, and incentives between public governments as well as including quality indicators in the efficiency measure.

Figure 1. Health expenditure 1995 and 2010 (public and private, % of GDP)

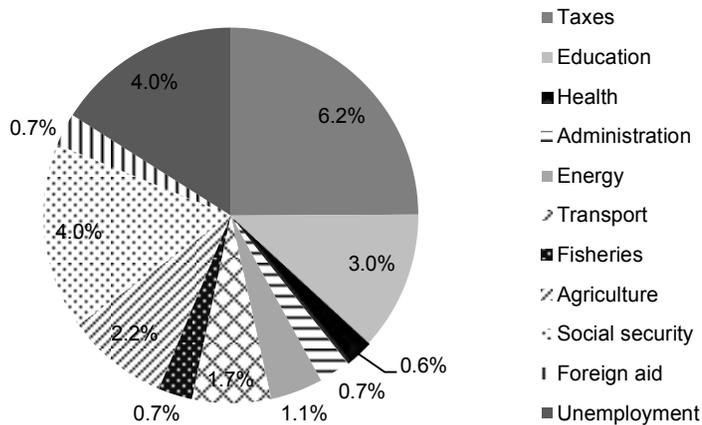


Source: World Bank, World Development Indicators (2011).

As the public sector grows larger and larger to meet the demands of the people for more and better public services, the need for efficiency in public expenditure as well as in tax collection becomes all the more urgent. This need is particularly urgent in the fields of education and health care because that is where most of the money is. The Nordic countries spend around ten percent of their GDP on health care provision, a figure

that could easily double in view of developments in the United States where public and private expenditures on health care approach twenty percent of GDP (Figure 1). Rehnberg and Häkkinen’s paper shows the way by suggesting how taxpayer money as well as private funds can be used more efficiently by producing more and better hospital services at a lower cost.

Figure 2. Benchmarking in Norway 1991: How to increase GDP by 25%



Source: Norman et al. (1991).

Even so, a seminal Norwegian study (Norman et al., 1991) suggests that the efficiency gains to be expected from benchmarking in the Norwegian hospital system are rather small as compared to the education system or the public sector as a whole (Figure 2). Using essentially the same method as Rehnberg and Häkkinen, the study by Victor Norman and his associates concluded that a frontal attack against inefficiency throughout the public sector could produce economic gains in the order of twenty-five percent of GDP. In the education system, for example, incentives could be put in place to encourage university students to graduate as young as they were when they graduated from university twenty years earlier, saving university resources and enabling the graduates to enter the labor market earlier. This recommendation was based on the fact that in 1990, it took a university student in Norway a year and a half longer on average to finish his or her studies than in 1970. More importantly, Nor-

man and his associates also advocated advancing the start of compulsory education from the age of seven to the age of six to add a year to the working life of each Norwegian. The study reported that the potential efficiency gains in education, all things considered, were five times as large as the potential efficiency gains in the health care sector, including the gains from inefficient hospitals emulating the methods applied by the most efficient hospitals as in Rehnberg and Häkkinen's analysis.

This result does not, however, diminish the importance of benchmarking in the hospital sector. Economists constantly need to look for ways of increasing efficiency and lifting the standard of life. This, simply put, is what economics is all about. There seems to be a significant scope for intensive economic growth in the Nordic countries with their large public sectors and in Europe, driven by a more efficient use of existing capital and other resources in contradistinction to extensive growth driven by the buildup of capital. Chinese school children now study English from the age of six.

References

- Norman, V. et al. (1991), *Mot bedre vitende? Effektiviseringsmuligheter i offentlig virksomhet*, Arbeids- og administrasjonsdepartementet, Bergen.
- World Bank (2011), *World Development Indicators*, World Bank, Washington, DC.



norden

Nordic Council of Ministers

Ved Stranden 18
DK-1061 Copenhagen K
www.norden.org

The Nordic Economic Policy Review is published by the Nordic Council of Ministers and addresses policy issues in a way that is useful for informed non-specialists as well as for professional economists. All articles are commissioned from leading professional economists and are subject to peer review prior to publication.

The review appears twice a year. It is published electronically on the website of the Nordic Council of Ministers: www.norden.org/en. On that website, you can also order paper copies of the Review (enter the name of the Review in the search field, and you will find all the information you need).



TemaNord 2013:514
ISBN 978-92-893-2496-0
ISSN 1904-4526