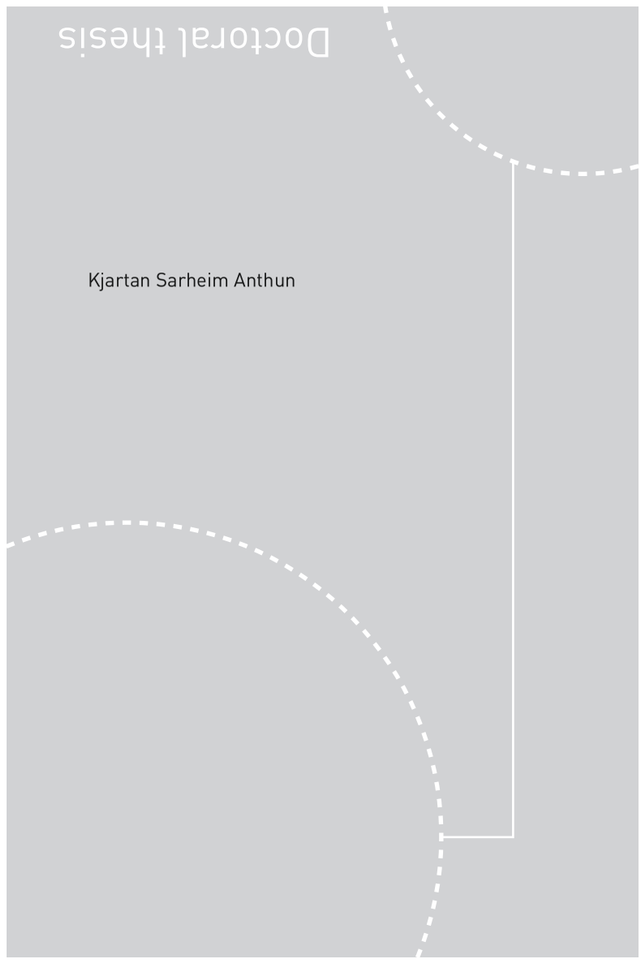
Doctoral theses at NTNU, 2017:302

Kjartan Sarheim Anthun

Productivity in the Norwegian hospital sector

Financing, quality, coding and comparability



ISBN 978-82-326-2672-4 (printed ver.)
ISBN 978-82-326-2673-1 (electronic ver.)
ISSN 1503-8181

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Medicine and Health Sciences
Department of Public Health and Nursing

Doctoral theses at NTNU, 2017:302

NTNU
Norwegian University of
Science and Technology

Kjartan Sarheim Anthun

Productivity in the Norwegian hospital sector

Financing, quality, coding and comparability

Thesis for the Degree of Philosophiae Doctor

Trondheim, October 2017

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Public Health and Nursing

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Medicine and Health Sciences
Department of Public Health and Nursing

ISBN 978-82-326-2672-4 (printed ver.)
ISBN 978-82-326-2673-1 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2017:302

Printed by NTNU Grafisk senter

Norsk sammendrag

Produktivitet blant norske sykehus: Finansiering, kvalitet, koding og internasjonal sammenlignbarhet

I en tid med aldrende befolkning og teknologisk utvikling er sykehustjenester i offentlige og universelle helsesystemer under stadig sterkere press. Sykehusforbruket til eldre er høyt, og forventninger om medisinsk og teknologisk fremskritt vil øke kostnadene ytterligere. Formålet med denne avhandlingen er å se på utviklingen i sykehusproduktivitet i perioden 1999 til 2014, og hvordan denne utviklingen har vært i forhold til måten sykehusene er finansiert på. Avhandlingen består av fire studier som belyser ulike sider av dette: 1) produktivitet og produktivitetsendringer i perioden 1999 til 2014, 2) komparativ analyse av produktivitetsendring i nordiske land, 3) forholdet mellom kvalitet og produktivitet og 4) forholdet mellom finansiering og diagnosekoding.

Studie 1 viser at produktiviteten til norske sykehus hadde en samlet gjennomsnittlig vekst på 24.6 prosent fra 1999 til 2014, eller en årlig vekst på 1.5 prosent. Den største produktivitetsveksten skjedde i årene rundt helseforetaksreformen i 2002. Etter sykehusreformen har de fleste sykehus blitt større enn det vi estimerer som optimal størrelse.

Videre sammenligner vi produktivitet blant sykehus i Norge, Sverige, Danmark og Finland. I studie 2 fant vi at gjennomsnittsproduktiviteten blant norske sykehus kun var 56.6 prosent sammenlignet med de mest produktive i Norden, som alle var finske. Dersom vi kun sammenligner norske sykehus med andre norske sykehus er den (tekniske) effektiviteten imidlertid like høy innad i Norge som innad i Finland. En mulig forklaring på ulik produktivitet er kvalitetsforskjeller. I studie 3 finner vi store forskjeller mellom sykehus når det gjelder både reinnleggelsesrater og mortalitet. Norske sykehus har generelt høyere reinnleggelsesrater enn de andre nordiske landene, men lavest mortalitetsrater. Sammenhengen mellom kvalitet og produktivitet er ikke entydig.

En alternativ forklaring på produktivitetsforbedringene er at sykehusene har blitt bedre til å dokumentere diagnosene til pasientene. Dette kan føre til at sammenligninger over tid vil overdrive utviklingen av økt produksjon. Vi finner i studie 4 at det er en viss sammenheng mellom prisinsentiv og bruk av kompliserende diagnosekoder. Imidlertid er effekten av prisendring langt mindre.

Studiene i denne avhandlingen er alle basert på analyser av data fra Norsk pasientregister i tillegg til data innsamlet fra sykehusregnskap. Statistiske metoder velegnet for store datasett har blitt benyttet.

Navn kandidat: Kjartan Sarheim Anthun

Institutt: Institutt for Samfunnsmedisin og Sykepleie

Veileder(e): Jon Magnussen og Johan Håkon Bjørngaard

Finansieringskilde: Norges forskningsråd

Ovennevnte avhandling er funnet verdig til å forsvares offentlig for graden Ph.D. i Samfunnsmedisin

Disputas finner sted i Auditoriet ØHA11, Øya Helsehus, fredag 27. oktober kl 12 15.

English summary

Productivity in the Norwegian hospital sector: Financing, quality, coding and comparability

In a time of aging populations and technology improvements, hospital services in universal public systems are under increasing pressure. Hospital utilization rates are also relatively greater for elderly, and expectations of medical and technological improvements will continue to increase costs.

The aim of this thesis is to identify developments in hospital productivity and its relation to the funding of the hospital sector in Norway during the period 1999 to 2014. The thesis contains four studies: 1) hospital productivity and productivity changes during the period 1999 to 2014, 2) a comparative analysis of hospital productivity growth in Nordic countries, 3) the relationship between hospital quality and productivity, and 4) the relationship between hospital financing and diagnostic coding.

In study 1 we found that the productivity of Norwegian hospitals had an average weighted growth of 24.6% from 1999 to 2014, or an annual increase of 1.5%. The largest gains occurred around the implementation of the hospital ownership reform in 2002. After the hospital reform, most hospitals are larger than what we estimate as optimal size.

Further we compare productivity amongst hospitals in Norway, Sweden, Denmark and Finland. In study 2 we estimated the mean productivity in Norway to be 56.6 per cent compared to the best Nordic hospitals, all being Finnish. If we compare productivity only within each country, the (technical) efficiency in Norway is as high as in Finland. A possible explanation of productivity differences is quality differences. In study 3 we found large differences between hospitals regarding both readmission rates and mortality. Norway had higher readmission rates than the other countries, but the lowest mortality rates. No clear cost–quality trade-off pattern was revealed.

An alternative explanation of productivity growth is that hospitals have changed diagnostic coding practices. This would exaggerate the measurable productivity growth over time. Study 4 show that there is an association between price incentive and the use of complicated diagnoses. However, the effect of price changes is smaller.

The studies in this these are all based on register data from the Norwegian Patient Registry, as well as hospital cost accounting data. Statistical methods suited for large dataset have been utilized.

Table of Contents

List of figures

List of tables

Acknowledgements

This thesis is the product of a PhD grant funded by the project "*The effects of DRG-based financing on hospital performance: productivity, quality and patient selection*" through the Health and Care Services Programme at the Norwegian Research Council (grant 214338/H10). *Beregningsutvalget for Spesialisthelsetjenesten* has funded part of the work done in Paper II, and Paper III is in part funded by the *European Union*'s 7^{th} *Framework Programme for Research and Technological Framework* (grant 241721). Finally, my employer SINTEF has also funded part of the time I have spent on this thesis.

My eager supervisors have been professor Jon Magnussen and professor Johan Håkon Bjørngaard from the Department of Public Health and Nursing at the Norwegian University of Science and Technology (NTNU). Each supervisor has shared valuable and different perspectives on my work. Both have been patient with my, at times, slow progress. I am similarly hugely indebted to Sverre Kittelsen from The Ragnar Frisch Centre for Economic Research, who has not been a supervisor but nevertheless has patiently tried to teach me, a non-economist, the workings of productivity studies.

Thank you also to my many co-authors. Two of the papers are a product of an ongoing Nordic collaboration, the Nordic Hospital Comparison Study Group, to which I am thankful to be a part of. Fanny Goude, Øyvind Hope, Ingrid M S Huitfeldt, Unto Häkkinen, Birgitte Kalseth, Jannie Kilsmark, Sverre A C Kittelsen, Marie Kruse, Emma Medin, Clas Rehnberg, Hanna Rättö and Benny Adam Winsnes have contributed to the great collaborative work in Paper II and Paper III. The other two papers have been written in collaboration with Sverre Kittelsen (Paper I), Johan Håkon Bjørngaard (Paper IV) and Jon Magnussen (Paper I and Paper IV). Both I and the final papers have benefited hugely from these experienced researchers' knowledge, wisdom and wit.

In addition to funding partners and co-authors I also wish to thank:

- The Norwegian University of Science and Technology, the Faculty of Medicine and the Department of Public Health and Nursing for employing me as a PhD student.

- While working (part time) as a PhD student I have also been employed as a research scientist at SINTEF. I have been fortunate to been allowed the time and space needed to work on this thesis, instead of solely working on more profitable projects. I especially benefited from this goodwill when the PhD deadline passed.
- Thank you to all Health Services Research colleagues in SINTEF and NTNU for commenting on early iterations of papers I have presented at in-house workshops.
- Tony Scott and The Melbourne Institute for Applied Economics and Social Research for inviting me to Melbourne. The visit regretfully only lasted for half a year, however the time visiting the University of Melbourne was the highlight of the time working on the thesis since the whole family of five went on this exciting journey.
- Participants at various workshops and conferences that have contributed to the research by discussing and commenting on different versions of my papers. A special mention goes to the participants at the Linguaglossa Colloquium who gave generous comments despite belonging to a different discipline of science.
- My children Jon, Tyra and Waldemar for constantly reminding me that life has other things to offer than just research, and at the same time accepting me as their quirky researcher dad.
- And finally, Kirsti, my extremely patient and supportive wife. From the beginning, you have been my most important academic inspiration. Throughout this work you have been the fervent organizer of the family and most importantly my best friend. My gratitude and love for you is endless.

Science proceeds by small incremental steps, and I hereby present my miniscule contributions. While standing on the shoulders of these mentioned giants, all errs presented here are my own.

Trondheim, June 2017

List of papers

Paper I: Productivity growth, case mix and optimal size of hospitals. A 16-year study of the Norwegian hospital sector

Kjartan Sarheim Anthun, Sverre A C Kittelsen, Jon Magnussen

Health Policy, 121(2017), 418-425

https://doi.org/10.1016/j.healthpol.2017.01.006

Paper II: Decomposing the productivity differences between hospitals in the Nordic countries

Sverre A C Kittelsen, Benny Adam Winsnes, Kjartan S Anthun, Fanny Goude, Øyvind Hope, Unto Häkkinen, Birgitte Kalseth, Jannie Kilsmark, Emma Medin, Clas Rehnberg, Hanna Rättö

Journal of Productivity Analysis, 43:281-293, 2015

https://doi.org/10.1007/s11123-015-0437-z

Paper III: Costs and Quality at the Hospital Level in the Nordic Countries

Sverre A C Kittelsen, Kjartan S Anthun, Fanny Goude, Ingrid M S Huitfeldt, Unto Häkkinen, Marie Kruse, Emma Medin, Clas Rehnberg, Hanna Rättö

Health Economics 24(Suppl. 2): 140-163 (2015)

https://doi.org/10.1002/hec.3260

Paper IV: Economic incentives and diagnostic coding in a public health care system

Kjartan Sarheim Anthun, Johan H Bjørngaard, Jon Magnussen

International Journal of Health Economics and Management, 2017 17:83-101

https://doi.org/10.1007/s10754-016-9201-9

1 Introduction

In a time of aging populations and technology improvements, hospital services in universal public systems are under increasing pressure. The foundation of Norwegian hospital policy is the National Health and Hospital Plan, which estimates growth in the number of elderly resulting in a 27 per cent increase in demand for hospital personnel by 2030 (Stortingsforhandling 2015). Hospital utilization rates are also relatively greater for elderly, so the aging population will only further increase the pressure on hospitals. While the expectation is that medical and technological improvements will continue to reduce the time patients spend in hospitals, it will likely be at an increased cost.

Further, in 1997, an activity-based financing (ABF) system was introduced in Norway, changing the way hospital services were financed from a retrospective global budget to a partly variable prospective payment system (PPS). One of the primary goals was increased productivity (Stortingsforhandling 1996); therefore, it is highly relevant to look at the details of how this financing scheme affects productivity and hospital production. Moreover, the ownership of Norwegian hospitals transferred from counties to the central government in 2002, forming an ownership structure consisting of semi-autonomous regional health authorities (RHAs) and hospital trusts.

The aim of this thesis is to identify developments in hospital productivity and its relation to ABF in the hospital sector in Norway during the period 1999 to 2014. Specifically, I examine four main issues: a) hospital productivity and productivity changes during the period 1999 to 2014, b) a comparative analysis of hospital productivity growth in Nordic countries, c) the relationship between hospital quality and productivity, and d) the relationship between hospital financing and diagnostic coding.

1.1 Outline of the thesis

Chapter 2 details the Norwegian health care system, which is the context of this thesis. I present some of the recent reforms and describe the funding scheme for specialized somatic hospital care. Chapter 3 describes the economic framework upon which this thesis builds.

Chapter 4 presents a literature review in which I focus on four different topics. First, I describe how hospital productivity and productivity change have been measured in the literature. Second, I present cross-national comparisons and discuss previous comparisons of hospitals in different countries. Third, I discuss how health outcomes and quality indicators relate to hospital productivity. Lastly, I consider how issues of financing (incentives) and productivity relate to diagnostic coding practices.

Chapter 5 summarizes the aims of this thesis, while Chapter 6 discusses the application of data and methods to attain these aims. Chapter 7 summarizes the findings from the various papers, and Chapter 8 presents the overall results.

2 The Norwegian health care system

Norway is a sparsely populated country on the northern fringe of Europe. Following the discovery of oil in the North Sea in 1969, Norway's income and welfare increased dramatically. Norway is now a country rich in resources, especially offshore oil and gas, and is considered a wealthy egalitarian society with high levels of political and social trust (Delhey and Newton 2005). Norway has a high gross national income per capita and scores well on the UN Human Development Index.

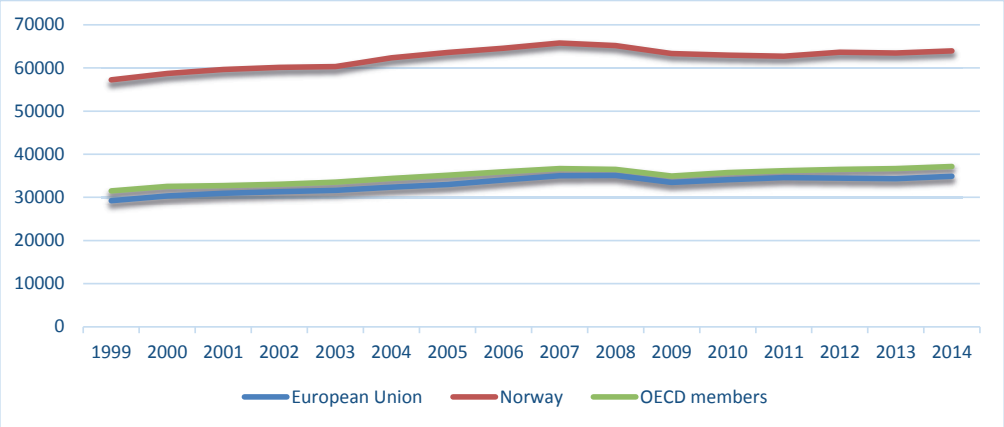


Figure 1 GDP per capita (PPP-adjusted constant 2011 international $), the European Union, Norway, and OECD member countries (source: World Bank, International Comparison Program database)

Figure 1 compares the developments in GDP per capita from 1999 to 2014 for Norway and Organisation for Economic Co-operation and Development (OECD) and European Union (EU) member countries. Adjusted for purchasing power parity (PPP), Norway was 82 per cent higher than other OECD countries in 1999, prior to which, the difference was stable. In 2014, Norwegian GDP per capita was 72 per cent higher than the weighted OECD average and 83 per cent higher than the EU average. However, the mean annual growth during this period in

Norway was less than one per cent, clearly demonstrating the impact of the 2008/09 global financial crisis.

By comparative standards, Norway is a rich country, and can therefore potentially spend substantial amounts on health. Figure 2 plots health expenditure per capita for Norway along with the means for the EU and OECD member countries. Currently, Norway ranks fifth-highest in the world in per capita health expenditure behind the U.S., Monaco, Luxembourg and Switzerland. However, while per capita expenditure is high, because of Norway's large GDP, the health expenditure share of GDP (9.7 per cent in 2014) is below the average in both the EU (10 per cent in 2014) and OECD countries (12.4 per cent in 2014), as depicted in Figure 3. As percentage of GDP, there has been some fluctuation in health expenditure in Norway, such that it was only marginally higher in 2014 than it was in 1999.

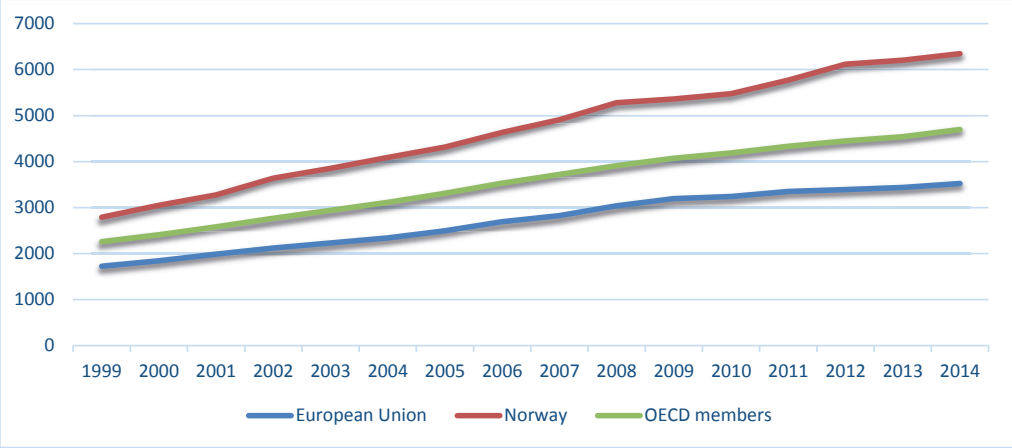


Figure 2 Health expenditure per capita (constant 2011 international PPP$) (source: World Health Organization Global Health Expenditure database)

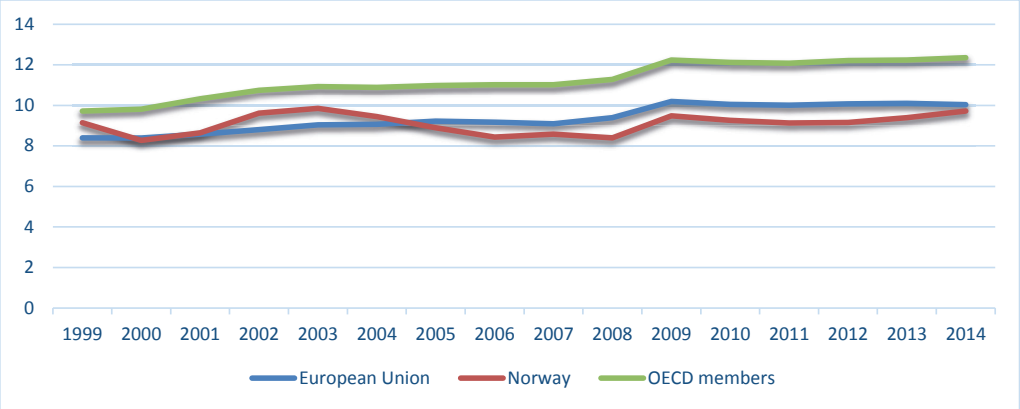


Figure 3 Health expenditure, total (% of GDP), the European Union, Norway, and OECD member countries (source: World Health Organization Global Health Expenditure database)

Norway has a National Health System (NHS)-type system mainly financed through taxation. Most public health services in Norway are free of charge at the point of use or require only a small out-of-pocket payment. While predominantly publicly funded, private actors (both for-profit and not-for-profit) provide some health services. For example, in primary care, most general practitioners are self-employed with contracts with municipalities. Of specialized services (hospitals), approximately 12.3 per cent of total costs were for private institutions in 2014 (Pedersen et al. 2016). Most of these costs relate to either institutions (*Avtaleinstitusjon*) or specialists (*Avtalespesialist*) with long-term public contracts that have not been subject to public tenders since before the 2002 hospital reform (Pedersen et al. 2016). Figure 4 illustrates the share of activity in each sector for hospital trusts, private institutions, and long-term contracted private actors. Of these, public institutions offer the most services, but in terms of rehabilitation and substance abuse, the public relies on private institutions for a large share of the total service provision.

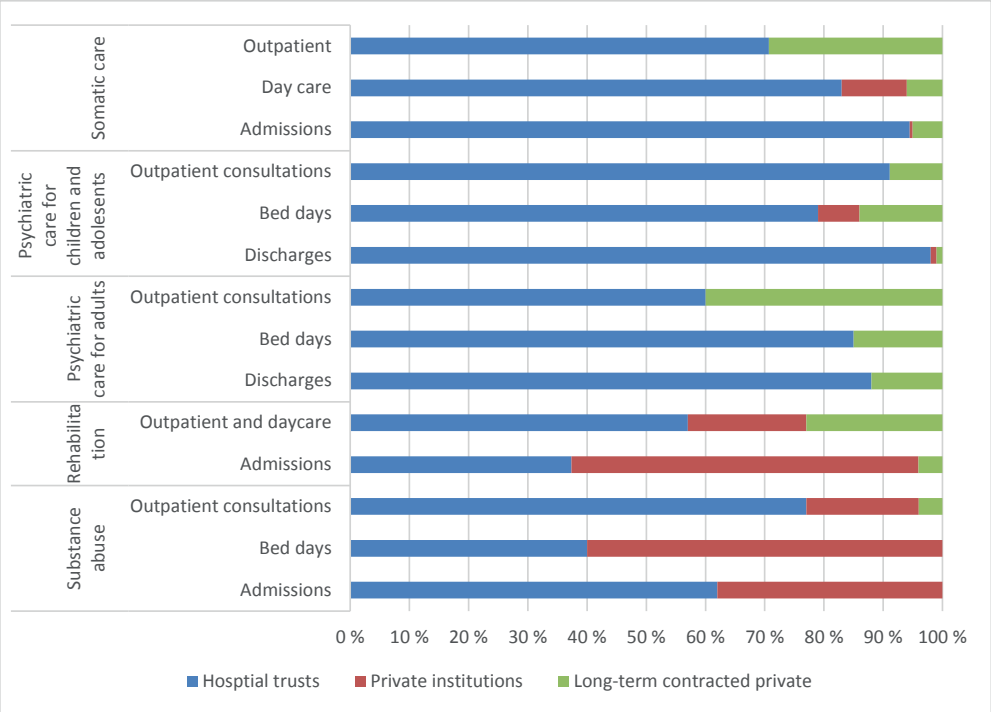


Figure 4 Share of activity by hospital trusts, private institutions, and long-term contracted private actors

Notes: Adapted using data in Pedersen et al. (2016). Somatic care activity is case-mix adjusted.

The responsibility for health services in Norway is at different governmental levels. Municipalities cover primary care (including general practitioners and long-term care), counties organize public dental care, and the central government is responsible for specialized secondary care, i.e., hospital services. This thesis studies productivity in acute care somatic hospitals, which account for more than 70 per cent of the total budget for secondary care (Samdata 2015).

2.1 The history and organization of Norwegian hospital services

The modern history of Norwegian hospitals began with the 1969 hospital law, in line with which, hospitals became the formal and legal responsibility of the 19 counties, with funding

shared by the counties and central government. At the time, this funding was mostly in the form of per diem transfers, which did not provide any incentive for hospitals to improve their productivity or undertake cost reductions. Consequently, 1980 saw its replacement by annual global budgets financed by the counties. In turn, the counties were financed partly by local taxation and otherwise by central government grants. The annual increases in expenditures were small, and questions were raised about hospital efficiency and long waiting lists (Biørn et al. 2003).

To change behaviour in hospitals, an ABF system based on diagnosis-related groups (DRGs) was implemented in 1997 (see Section 2.2 for details on hospital funding in Norway), following an initial trial in 1991/93 (Magnussen and Solstad 1994). The purpose of this funding change was to increase hospital activity, curb waiting lists and increase efficiency (Biørn et al. 2003; Stortingsforhandling 1996). Activity did indeed increase, but the funding led to budget gaming between counties and the central government (Tjerbo and Hagen 2009). Tjerbo and Hagen argued that "…agents expecting a soft budget constraint have incentives to increase activity or costs above what is preferred by the principal, send the bill to the principal and hope for bailouts," which was exactly what occurred after the introduction of ABF. The reimbursements from ABF were also lower than marginal costs, which only accentuated the budget deficits.

Following a period of increasing waiting times and budget deficits, in 2002, the Norwegian government transferred the ownership and control of hospitals from county councils to the central government (Hagen and Kaarbøe 2006; Magnussen et al. 2007). The main argument was to concentrate the responsibility of financing and ownership. This reform established the current structure of governance at three levels: the RHAs, hospital trusts and hospitals/institutions. The RHAs also own and operate trusts for hospital pharmacy enterprises, building, information technology, and supporting services.

According to *Spesialisthelsetjenesteloven* [*The Law of Specialized Hospital Care*], the RHAs are responsible for providing secondary care services for people in their health region. These secondary care services include hospital services (somatic, psychiatric and some rehabilitation), laboratory and radiology, emergency care, medical emergency call services, ambulance (plane, helicopter, car and boat), specialized treatment for substance abuse, transport to these services, and the transport of personnel.

Table 1 Number of public hospitals and hospital trusts, 1999–2014

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Hospitals	53	52	52													
Hospital trusts				34	28	28	27	27	26	26	21	20	20	20	20	20

Notes: Hospitals and hospital trusts included in the study, excluding private institutions. The actual number of hospitals and hospital trusts is higher as we exclude non-acute somatic care hospitals along with non-treatment hospital trusts.

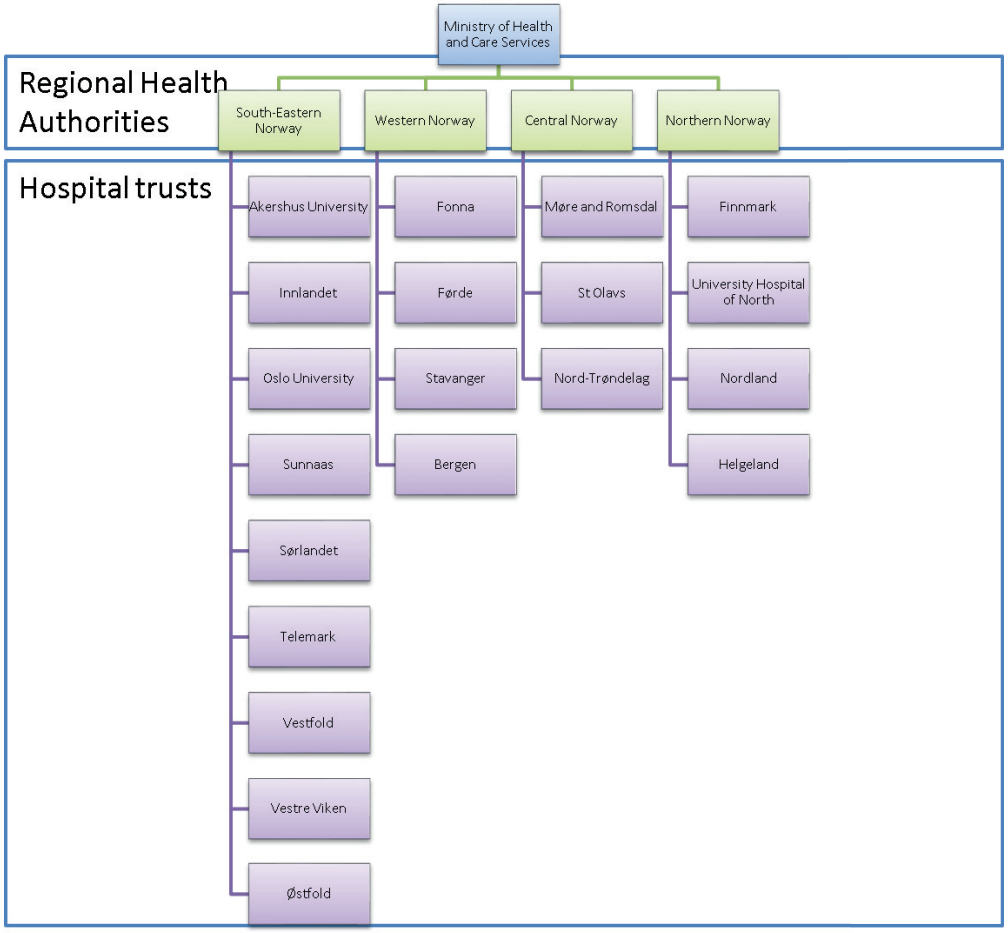


Figure 5 Organization of Norwegian hospitals, 2016

Before the 2002 reform, the Norwegian hospitals were owned by the 19 counties (except for some state-owned hospitals, among them, the National Hospital of Norway [*Rikshospitalet*]), summing to 55 public hospitals. The reform formed five RHAs and organized the hospitals into 43 hospital trusts. In the years following, many hospitals and hospital trusts reorganized and merged. In 2007, two of the RHAs merged, so now, each organize between three and nine hospital trusts (see Figure 5), currently totalling 20 hospital trusts, and each hospital trust may contain one or more hospitals. Table 1 details the number of public hospitals and public hospital trusts included in the study. As noted, some private non-profit hospitals have long-term contracts with an RHA, and are thus fully funded by the public to produce health services; two such hospitals are included in this study.

The final two major reforms in Norway were the *Samhandlingsreformen* [Coordination Reform] of 2012 and *Fritt Behandlingsvalg* [Patient Right to Choose Treatment Centre] of 2015. The first of these reforms introduced policies aimed at further reducing the length of patient stay in hospital and involving municipalities through co-payments and the establishment of emergency care beds in municipalities. The second reform enabled patients with a referral to choose a treatment centre regardless of private/public status. Beyond this, these reforms not discussed in this thesis.

2.2 Hospital funding in Norway

ABF in Norway depends on DRGs, with each hospital episode grouped in a specific DRG based on patient age and sex, diagnoses and procedures and length of stay. There are currently approximately 780 different DRGs, each having a specific cost weight; the reimbursement of RHAs in the year *t* is based on the following equation:

$$ABF = ABFSHARE_t \times DRGPRICE_t \times \sum COSTWEIGHT_{DRG,t} \times EPISODES_{DRG,t} \quad (1)$$

where ABFSHARE is the yearly reimbursement share for each episode, currently 50 per cent (see Figure 6) and DRGPRICE is the monetary value of each DRG point (i.e., COSTWEIGHT

= 1). In 2014, the DRGPRICE was 40,772 NOK. The COSTWEIGHT varies for each DRG and is a relative weighting system calibrated so that the average treatment has a value of one. EPISODES are the number of hospital episodes (including inpatient discharges, day care treatments/surgeries and outpatient treatments/consultations) in the region. It is optional for RHAs to continue this redistribution *within* their region, but now all regions do.

The expectation was that the use of DRGs for funding could invoke changes in diagnostic coding practices (Simborg 1981; Carter and Ginsburg 1985; Carter et al. 1990), so the Norwegian system was set up with a so-called "creep-ceiling" (Biørn et al. 2003; Stortingsforhandling 1996). Growth in the average treatment severity/intensity (case-mix index) beyond this creep-ceiling would not be funded, and thus this was also an instrument intended to contain costs (Kjerstad 2003). This ceiling likely limited the changes in diagnostic coding practices in the early years after 1997 (Biørn et al. 2003), but was removed in 2002.

At the time of the introduction of ABF, it was argued that too low of an ABF share would not have any effect, while too high of a share could lead to excessively strong incentives and cause unwanted shifts in production (Stortingsforhandling 1996). Since 1997, shifting political priorities in the Norwegian parliament have frequently led to changes in the ABF share, as shown in Figure 6.

When introduced in 1997, ABF determined the reimbursement of inpatient treatment and day care only. In 2009, the structure of DRGs changed with the creation of specific groups for day care (O-groups), and outpatients were included in the grouping and financing. The ABF share is how large a share of each treatment will receive reimbursement, while a global budget covers the remainder. However, not all hospital activity is DRG-financed, so the total income generated by DRG activity is less than the share shown in Figure 6. The global budget is distributed to the RHAs based on risk-adjusted capitation.

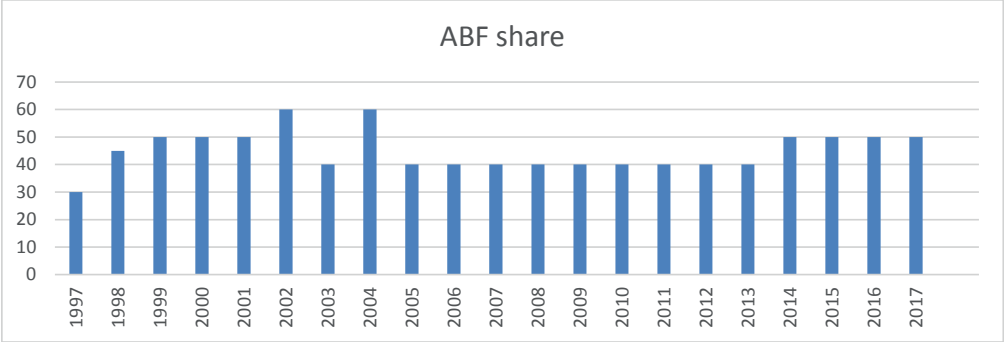


Figure 6 ABF share

The prices for each DRG are updated annually based on average costs two years prior for treatment in the same DRG (before 2005 the weights were updated bi-annually). A committee (*Avregningsutvalget [Settlement Committee]*) organized by the Directorate of Health is responsible for auditing the funding scheme and harmonizing the use of the medical coding related to ABF. The committee can reduce the total transfers to the RHAs if they identify examples of misuse.

3 Efficiency and productivity

Productivity is the conversion rate of production as measured by the ratio of inputs to outputs. Efficiency is a theoretical measure comparing the observed productivity to the best possible productivity. According to Jacobs et al. (2006), the entities studied in productivity analyses must: 1) capture the entire production process, 2) be decision-making units (DMU), i.e., have discretion about the conversion from inputs to outputs, and 3) be comparable. When referring to the economic literature, we use the term units, such that in the empirical part of this thesis, the units will be hospitals and hospital trusts.

The production possibility set is the possible combinations of inputs and outputs.

$$y = f(x) \qquad (2)$$
$$(x, y) \in T$$

where *y* is the output, *x* is the input, and *(x, y)* is the input-output combination that must belong to the production possibility set (or technology) *T*. If there is only one output *y*, the production function *f(x)* describes the frontier. Units that produce on the frontier are efficient. We assume that it is always possible to use more resources at a certain output level or produce less of a service for the same input level. The units not producing on the frontier are inefficient.

3.1 Graphical example: defining efficiency

Consider a hypothetical setting with one input and one output, and a proportional relationship between them, as illustrated in Figure 7. An example of this setting is the number of patients treated (the output) by the number of physicians (the input). The production possibility set is below and to the right of the line OC. Any efficient unit would produce on the line OC, and inefficient units would lie below OC. We let A, B_{C0} and B_{C1} be units.

Unit B_{C1} is on the production possibility frontier and uses X_1 inputs to produce Y_1 outputs. Its efficiency is how the distance X_1Bc_1 relates to the maximum output for the same input, which is also the distance X_1B_{C1}. Unit B_{C1} thus has efficiency $X_1B_{C1}/X_1B_{C1}=1$. Any output smaller than Y_1 for the same input X_1 is inefficient. Unit A produces Y_0 outputs for the same amount of input as unit B_{C1}. The productivity of unit B_{C1} determines the efficiency of unit A, being the distance X_1A divided by the distance X_1B_{C1}, i.e., X_1A/X_1B_{C1}. As $A<B_{C1}$ the efficiency of unit A is <1.

The measurement of productivity and efficiency can be in both input and output directions. In an input orientation, we measure how much the reduction of an input is possible while keeping output constant. The reverse is true in an output direction, that is, how much we can expand output without increasing the use of inputs. We assume units are then either input-minimizing or output-maximizing. The example above is output oriented as we compare units B_{C1} and A for the same input X_1. In the input direction, we would hold the output constant and compare A to B_{C0} on the frontier. The efficiency of B_{C0} is $Y_0B_{C0}/Y_0B_{C0}=1$. The efficiency of unit A in the input direction is $Y_0B_{C0}/Y_{C0}A$. Given $A>B_0$, it follows that $Y_0B_{C0}/Y_0A<1$.

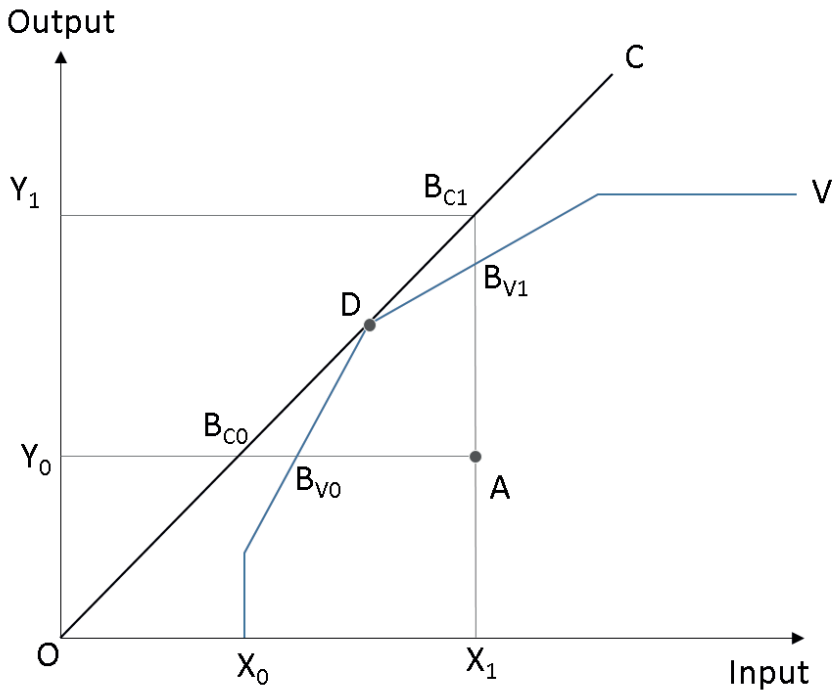


Figure 7 Production frontier and efficiency under constant and variable returns to scale (VRS)

3.2 Efficiency: technical, allocative and scale

We can distinguish between different forms of efficiency. Technical efficiency is the distance from a unit to the production possibility frontier with a similar mix of inputs and outputs. Allocative efficiency, also called price efficiency, reflects the ability of a unit "…to use inputs in optimal proportions given their respective prices" (Jacobs et al. 2006) and implies an optimal mix of different inputs. Total efficiency (sometimes called cost efficiency or economic efficiency) is the product of technical efficiency and allocative efficiency.

The final efficiency term we encounter is scale efficiency. We distinguish between constant returns to scale (CRS; proportionality in the conversion of inputs to outputs) and variable returns to scale (VRS; a differentiated conversion ratio depending on the size or scale). Optimal size is where average costs are minimized and CRS=VRS. Scale efficiency is also measured

either in an input or output direction and is the distance from the CRS to the VRS frontier at either the level of output or the level of input.

The line OC in Figure 7 is an example of a technology with CRS, comprising a linear (and thus constant) conversion from inputs into outputs. However, this relationship is often more complex, in that fixed costs may imply increasing returns to scale. However, at some point, increased costs of, for example, coordination, may lead to decreasing returns to scale. In this case, there is no proportionality between inputs and outputs; thus, we have VRS. Figure 7 depicts an example of a technology with VRS as the curve X_0DV.

If we consider a unit A, under CRS, we measure efficiency as the ratio Y_0B_{C0}/Y_0A in the input direction, while under VRS, efficiency is the ratio Y_0B_{V0}/Y_0A. Thus, under VRS, inefficient units will have a higher efficiency than under CRS, given $B_V > B_C$. Scale efficiency is a measure of the relative distance from B_C to B_V, and for unit A, scale efficiency (in the input direction) is measured by the ratio Y_0B_{C0}/Y_0B_{V0}. It follows that CRS efficiency is the same as VRS efficiency multiplied by scale efficiency. The scale efficiency is only dependent on the relative distance from B_{C0} to B_{V0} (or B_{C1} to B_{V1}) and not on the optimal scale, so scale efficiency may be high at the same time the distance to the optimal scale is large.

3.3 Productivity change and efficiency change

The terms so far define productivity and efficiency at a fixed point in time, but in this thesis, we examine productivity and efficiency change. A Malmquist index (Caves et al. 1982; Färe et al. 1994a) is defined as the productivity in time t+1 divided by the same unit's productivity in time t. This index will be >1 if there is productivity growth, and <1 if productivity falls. We can then decompose the Malmquist index as the product of three changes: frontier change, scale efficiency change and technical efficiency change.

Frontier change is the change in technology over time. That is, how the CRS frontier changes. Scale efficiency change is how the scale efficiency of a unit develops, and technical efficiency change is the pure efficiency change, that is, how the unit's distance to the current frontier changes (or catches up).

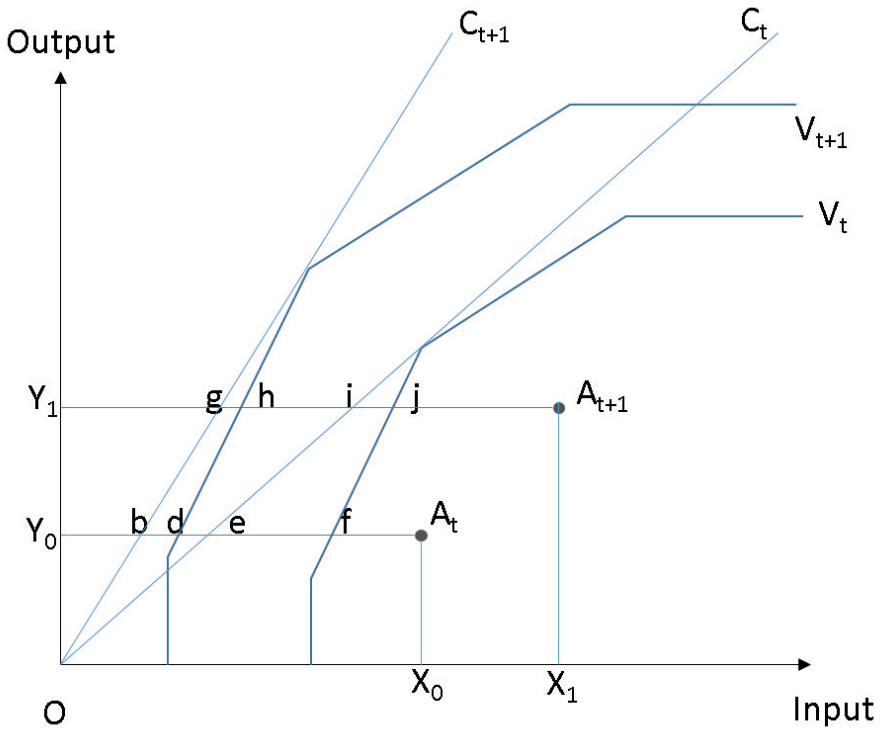


Figure 8 Illustration of productivity change

Figure 8 (adapted from Jacobs et al. (2006) illustrates the concept of productivity change, simplified here only in the input direction. We expand Figure 7 to two periods, time t and time t+1. As above, the input-oriented technical efficiency of A is the relationship between the distances Y_0f and Y_0A_t. The shift from time t to time t+1 implies for A_t increased input and output, and the Malmquist index is a measure of this change. However, we are interested in the decomposition because at this stage, three changes are taking place: 1) the frontier change, 2) the scale efficiency change and 3) the technical efficiency change (or the pure efficiency change). The Malmquist index is a product of these changes.

The change in technology (a frontier shift) is the change in the scale-efficient technology (i.e., measured relative to the CRS).

$$frontier\ shift = \frac{{}^{Y_0 e}/_{Y_0 A_t}}{{}^{Y_0 b}/_{Y_0 A_t}} = \frac{Y_0 e}{Y_0 b} \tag{3}$$

Pure efficiency change is the change in a hospital's distance from the current technically efficient frontier (under VRS) from time *t* to *t+1*.

$$technical\ efficiency\ change = \frac{\left(\frac{Y_1 h}{Y_1 A_{t+1}}\right)}{\left(\frac{Y_0 d}{Y_0 A_t}\right)} \tag{4}$$

Finally, scale efficiency is the change in scale efficiency from time *t* to time *t+1*

$$scale\ efficiency\ change = \frac{\left(\frac{Y_1 g}{Y_1 A_{t+1}} \Big/ \frac{Y_1 h}{Y_1 A_{t+1}}\right)}{\left(\frac{Y_0 e}{Y_0 A_t} \Big/ \frac{Y_0 f}{Y_0 A_t}\right)} = \frac{\left({}^{Y_1 g}/_{Y_1 h}\right)}{\left({}^{Y_0 e}/_{Y_0 f}\right)} \tag{5}$$

4 Current evidence

This chapter reviews the literature on hospital productivity. The organization of this chapter reflects the issues discussed in the thesis. In the health care sector, studies have been performed at different levels for a wide range of purposes by comparing countries (Puig-Junoy 1998; Varabyova and Schreyögg 2013), regions (Gerdtham et al. 1999; Halkos and Tzeremes 2011), specific services or departments (Hollingsworth and Parkin 2001), physician workforces (Johannessen et al. 2017) and even individual physicians (Chilingerian 1994). However, the most common has been at the hospital level (Hollingsworth 2003, 2008).

4.1 Case-mix comparison

Hospitals treat many types of patients, and their production is thus of a multiproduct nature. It is therefore vital to use some form of case-mix adjustment to weight the different outputs and ensure the comparison of similar products. In addition, as there are so many different types of patients, it is important to aggregate the outputs into meaningful and manageable categories for analysis. This section presents some of the studies on case-mix adjustment and discusses the issues of measuring change over time.

It is easiest to measure hospital output as simply the number of patients or the number of bed days; however, such operationalization underestimates differences in technology, the quality of care, case severity, and institutional characteristics (Tatchell 1983). Case-mix adjustment is a term that describes adjustment for known differences to improve output measures, and according to Averill (1991), this stems from at least five dimensions: severity of illness, prognosis, treatment difficulty, the need for intervention, and resource intensity. From an economic perspective, the resulting case-mix can be interpreted as the "…resource intensity demands that patients place on an institution" (Averill 1991). The importance of case-mix has been well documented in different parts of health services, including somatic hospitals (Barer 1982), long-term care (Berlowitz et al. 1996), child and adolescent mental health services

(Halsteinli et al. 2010), primary care (Salem-Schatz et al. 1994), and even when analysing infection rates (Sax et al. 2002).

Chowdhury et al. (2014) tested the importance of case-mix adjustment on a five-year panel of 113 hospitals in Ontario, Canada. They found that case-mix adjusting output was better than not adjusting, and better than including a simple case-mix as a separate output, given adjusting provided "…greater discriminatory power in distinguishing hospitals on efficiency and productivity meters" (Chowdhury et al. (2014):79). However, as a Danish study found when including additional patient characteristics, which did not yield any more statistical explanatory power than a simple case-mix index (Hvenegaard et al. 2009), merely increasing the number of controls may not be the optimal strategy. When conducting analysis with a limited number of observations (such as a national cross-section of hospitals), there is also a severe trade-off between the number of included variables and statistical power.

These studies show that adjusting for case-mix may be important, but there is no consensus on how it should be done other than some variation over using the DRG system to case-mix adjust the output. Further, the optimal aggregation strategy, either theoretically or empirically, remains unclear. Chowdhury et al. (2014) adjusted case-mix by weighting *individual level episodes* differently before aggregating. Grosskopf and Valdmanis (1993) adjusted *hospital level* data by average case-mix and found no significant difference between the case-mix adjusted and unadjusted results. To further this research, Magnussen (1996) analysed how the operationalization and aggregation of hospital outputs impacted efficiency measurement, with overall hospital efficiency found to be robust to the choice of aggregation of output, but not the rankings and scale results.

A Finnish study tested the impact of different levels of adjustments on measured efficiency (Vitikainen et al. 2009). The results did not differ much between episode- and admission-level data, as both of these are still case-mix adjusted. They also tested how DRG grouping only inpatients compared with DRG grouping both inpatient and outpatients. At the hospital level, there were differences between the two approaches. Vitikainen et al. (2009) suggested using a so-called full version of DRG grouping as opposed to inpatient grouping only to ensure a better case-mix adjustment of outpatient treatments.

Case-mix adjustment is the adjustment between units and hospitals. However, it is just as important to adjust for changes over time. Many studies and evaluations of reforms commonly

apply a two-year time span, one year before and one year after a given reform. Most of the studies reviewed by O'Neill et al. (2008) did not cover multiple years. Similar newer results are also presented by Chowdhury et al. (2014), wherein most of the studies utilize either a one-year cross-sectional analysis or a two-year comparison. However, to establish the long-term effects of reforms and policy changes, it is important to use data over a longer time span. This has some obvious difficulties related to data availability, and just as importantly, data comparability. Over time, different hospitals will treat different patients and experience different technologies.

Obviously, not all studies span just one or two years, but hospital-level studies conducted from a long-term perspective are uncommon. Most long-term studies apparently offer no longitudinal case-mix adjustment (e.g., (Chowdhury et al. 2014; Rutledge et al. 1995; Sommersguter-Reichmann 2000)). There are only two Norwegian studies adopting a longitudinal perspective (Biørn et al. 2003; Halsteinli et al. 2010), and while these studies case-mix adjusted within each year, neither applied any adjustment for comparing output measures over the nine or 10 years in their respective analyses.

Another approach when analysing in a longitudinal perspective is to use a comparative control group by comparing to trends in other countries. This is done in Kittelsen et al. (2008) when comparing Norway to other major Nordic countries. They used common fixed cost weights to allow for comparisons across countries and over time. If not used, there would be a chance that different weights across the years would capture the technical change (frontier shifts). Section 4.4 below further elaborates upon the comparative approach.

These studies clearly demonstrate the need for proper case-mix adjustment. A longitudinal perspective might provide better evidence of policy results and a better estimate of the development. In this thesis, we expand the literature through increasing longitudinal comparability by ensuring that outputs and inputs are comparable over a longer time.

4.2 Decomposing productivity development

Productivity changes are a combination of technical change (the expansion of the productivity frontier, i.e., the best units improving) and pure efficiency change (inefficient units catching up behind the frontier). It is then possible to decompose productivity change as the product of

technical change and pure efficiency change, i.e., Malmquist decomposition (Färe et al. 1992, 1994a; Färe et al. 1994b). In one review (Hollingsworth 2008), most studies were cross-sectional in nature; only 8 per cent of these used the Malmquist approach (see Section 3.3 above). Hollingsworth suggests that the absence of commercial off-the-shelf software for Malmquist studies has so far limited the number of studies (Hollingsworth 2008). This could indicate that research in this field is not yet exhausted.

Maniadakis and Thanassoulis (2004) calculated Malmquist indices in a setting where producers are input-minimizing and prices are known. They claim that this makes "…it possible to identify the root sources of productivity changes". However, it also relies on allocative efficiency and requires input prices, which are not known in our empirical setting (see Section 6.1 below for details of the available data). Three other studies found technological change, but no pure efficiency change (Burgess Jr and Wilson 1995; Chowdhury et al. 2011; Sommersguter-Reichmann 2000), and none of these had a long-term perspective.

The few studies that exist indicate that research in this field is not exhausted, and we contribute to this literature in at least two ways: first, by decomposing a long time series instead of just a two-year period, and second, by decomposing the national frontier for use in comparative studies.

4.3 Optimal scale

In purely economic terms, firms will produce at the point where long-term average costs are at a minimum. However, for hospitals, especially for public hospitals, which usually do not have the discretion to decide their own scale, this might be difficult to maintain. Consequently, proposed hospital mergers often raise the issue of optimal scale, when the size of (public) hospitals may change. There are studies on hospital mergers that also estimate their impact on efficiency (i.e., Kjekshus and Hagen (2007)), and some of these also estimate the scale properties of hospitals (Given 1996; Dranove 1998; Posnett 1999; Ahgren 2008; Marini and Miraldo 2009; T. Kristensen et al. 2010).

Optimal hospital size varies between countries and across health systems. When studying hospitals in Ontario, Canada, Preyra and Pink found that while the optimal hospital size was

179 beds, the statistical models were not optimal for special or very large hospitals (>500 beds) (Preyra and Pink 2006). Hsing and Bond found that the highest productivity among U.S. hospitals was for hospitals with 272 beds (Hsing and Bond 1995). T. Kristensen et al. (2008) reported that there is some evidence in the literature for economies of scale for hospitals <200 beds: the optimal scale seems to be around 200–400 beds, and the average cost increases in the range of 400–600 beds. T. Kristensen et al. (2008) empirically found 275 beds to be the optimal size for Danish hospitals, while a later publication using the same sample asserted that the optimal size was approximately 350 beds (T. Kristensen et al. 2012). Johannessen et al. (2017) recently estimated optimal hospital size in Norway as approximately 350 beds based on data for 2001–2013. Only one of these four studies provide a confidence interval (CI) of the optimal size estimation, with T. Kristensen et al. (2008) showing this to be very large (180–585 beds) at the 95 per cent level.

Unsurprisingly, the concept of optimal scale has attracted some criticism. Magnussen claimed that optimal scale is a theoretical measure that compares observations to some unobtainable productivity (Magnussen 1996). A Danish study found that the results of scale estimates were highly dependent on the functional form of the economic model (T. Kristensen et al. 2008). Likewise, Dranove warns that hospital size too often is expressed in terms of inputs rather than outputs, and claims that this "…can lead to misleading conclusions about the magnitude of scale economies if, as one suspects, hospitals with higher levels of output use their capacity more efficiently" (Dranove 1998). Similarly, T. Kristensen et al. (2012) also warns against using the number of beds as a proxy for costs. Aletras (1999) suggests that more scale studies should use short-run instead of long-run cost functions, as the latter tends to favour the presence of economies of scale. Asmild et al. (2013) found that optimal scale was not consistent across hospitals in different locations. Lastly, Førsund and Hjalmarsson (2004) show that the data envelopment analysis (DEA) method may yield results that indicate optimal scale for very different combinations of inputs and outputs, an argument that implies caution with respect to policy recommendations about scale efficiency.

In our analysis, we estimate optimal hospital size based on the results for productivity and scale efficiency, and calculate optimal size not as the number of beds, but as costs. Dranove recommends using outputs, but given the multiproduct nature of hospitals, it is more convenient to use costs.

4.4 Comparative studies of productivity

In Norway, national benchmarking reports are published annually (Samdata (2015)). However, the increasing flow of policies across borders adds value to comparative studies, especially in geographically close and economically and socially similar countries such as the Nordic countries. Cross-country comparisons serve two additional important concerns. First, hospital units from other countries can serve as controls for concurrent trends. Second, using observations from more than one country may improve statistical strength by increasing the number of observations. Using units from other countries may also improve our knowledge about variations in health policy and the role of institutional differences.

A few studies have used aggregate statistics to describe health system performance (Färe et al. 1997; Puig-Junoy 1998; Retzlaff-Roberts et al. 2004; Varabyova and Schreyögg 2013). These studies generally rely on national level databases to describe the overall performance of health systems, but do not disclose any adjustments to increase or ensure comparability. One study reports the inclusion of older data for missing data (Retzlaff-Roberts et al. (2004)). Results vary; for instance, Norway has been measured as being highly productive (Varabyova and Schreyögg 2013) or on the frontier (Retzlaff-Roberts et al. 2004) in some studies, whereas another study found that the Nordic countries were the most inefficient, and that this inefficiency was "…associated primarily with non-increasing returns to scale" (Puig-Junoy 1998). Aggregate studies such as these have been found to be particularly sensitive to the type of estimation and the available data (Hollingsworth and Wildman 2003; Spinks and Hollingsworth 2009).

A major issue with comparative studies is that the basic unit of analysis, hospitals, differs widely between otherwise similar countries. The delineation of different services and the content and intensity of hospital services differ; thus, the scope of hospitals and environmental settings are not easily comparable across countries, and the number of cross-country comparative studies on productivity has been limited (Jacobs et al. 2006). Most early studies comparing productivity at the hospital level have compared two countries, and the measures were mostly unadjusted, with comparability approximated by selecting the most (presumably) comparable units, inputs and outputs (Dervaux et al. 2004; Linna et al. 2006; Mobley and Magnussen 1998; Steinmann et al. 2004). For instance, in their comparison of France and the

U.S., Dervaux et al. (2004) found that it was not possible to estimate productivity using the same technology (frontier).

Increased EU funding for comparative research has expanded the number of European studies in recent years (see e.g., Street et al. (2011); Busse (2012); Aiken et al. (2014); Häkkinen et al. (2015). Some of these European studies have included Norway, but for Norway, the Nordic countries offer the most interesting and relevant comparisons. Even though there are dissimilarities, the Nordic countries share common traits of having open economies, being sparsely populated, and having high levels of trust and taxation (Lyttkens et al. 2016). When comparing Norway, Sweden, Finland, and Denmark, Kittelsen et al. (2008) created price indices to address differences in prices, inflation and currencies, and outputs were adjusted by a common set of weights along with some adjustments in the DRGs to increase comparability. A similar approach was used also by Medin et al. (2011) and Linna et al. (2010). These Nordic studies showed that hospital productivity in Norway was not as high as that in Finland and Denmark, but higher than that in Sweden (Linna et al. 2006; Kittelsen et al. 2008; Linna et al. 2010; Medin et al. 2011; Medin et al. 2013). However, the differences between the countries have not yet been fully explored to determine what makes some hospitals in some countries more productive than in others. More work should be done on increasing the comparability between countries even further.

4.5 Health outcomes, quality indicators and efficiency

The purpose of health care is to improve health, but as real health outcomes are inherently difficult to measure, analyses of productivity and efficiency tend to focus on health care services rather than health outcomes. However, these "…measures are manifestly inadequate, as they fail to capture variations in the effectiveness (or quality) of the health care delivered" (Jacobs et al. 2006). In the past decade, a few studies have questioned how quality relates to hospital efficiency. Different approaches have been attempted, as some have analysed the potential trade-off between quality, health outcomes and efficiency, while others have tried to measure quality as part of hospital output.

Nayar and Ozcan tested how the inclusion of quality measures affected efficiency estimates, and found that among 53 U.S. hospitals, "…quality outcomes were not being compromised by the efficient hospitals" (Nayar and Ozcan 2008). McKay and Deily (2008) analysed U.S. hospitals using stochastic frontier analysis (SFA) to measure mortality and complications as a function of cost inefficiency. They found no trade-off between efficiency and quality, as the results were the same regardless of whether the hospital output included outcomes. In a large cross-sectional study of 1,377 hospitals, DEA was used to assess the quality/efficiency trade-off (Valdmanis et al. 2008). That study found that only three per cent of the total inefficiency could be attributed to quality congestion, and suggested that there was no trade-off between quality and efficiency (Valdmanis et al. 2008). Another U.S. study found no trade-off between quality and efficiency, but noted that it would be important to include more than one quality measure to avoid missing "…opportunities for performance improvement" (Clement et al. 2008). A review of 61 U.S. studies yielded mixed results for the relationship between quality and cost of care (Hussey et al. 2013).

Two Danish studies considered mortality and complications related to costs, and found that the ranking of hospitals differed depending on the inclusion of quality as an output (Hvenegaard et al. 2011; Kruse and Christensen 2013). Ferrier and Trivitt (2013) compared DEA with quality indices to control for quality. They found that controlling for quality significantly altered the efficiency results, and that "…adding quality as additional output accounts for some of the previously observed inefficiency" (Ferrier and Trivitt 2013). A British study attempted to measure outcomes differently by linking hospital costs to patient-reported outcomes (Gutacker et al. 2013). However, there were only very small effects from hospital costs to outcome, and only for one of the four procedures considered in their analysis.

To further improve quality and performance, quality has increasingly been encouraged by payment schemes such as pay-for-performance (Eijkenaar et al. 2013; Ogundeji et al. 2016; Or and Häkkinen 2011; Milstein and Schreyoegg 2016), and in some cases, even penalties for non-performance (S. R. Kristensen 2016). However, none of these studies has tested the relationship between health outcomes, quality indicators and efficiency from a comparative perspective. Is there more variation within or between countries, and could the differences in health outcomes or quality indicators explain some of the country-level differences described in Section 4.4 above?

4.6 Financing, incentives and coding

Medicare is a federal U.S. social insurance scheme covering the population aged 65 years and older; it also includes a few specific diseases for younger people. In the early 1980s, Medicare funding changed from a retrospective cost-based system to a PPS using DRGs to reimburse each hospital for each hospital stay. Simborg warned against this system, claiming that it would cause a "…deliberate and systematic shift in a hospital's reported case-mix in order to improve reimbursement" and referred to the phenomena as "upcoding" (Simborg 1981). This prediction happened as the average case-mix increased following the introduction of Medicare in 1983 (Steinwald and Dummit 1989; Carter and Ginsburg 1985; Carter et al. 1990; Ellis and McGuire 1986; Rosenberg and Browne 2001; Stern and Epstein 1985). Nevertheless, the effectiveness of the Medicare PPS led to the adaptation of DRGs for funding in several other countries (Geissler et al. 2011).

The use of DRGs in funding provides a strong incentive for enhancing efficiency, but the results have been mixed (Street et al. 2011). A Cochrane overview of reviews on the effectiveness of financial incentives found mixed results depending on the type of financial incentive, but no evidence that financial incentives improved patient outcomes (Flodgren et al. 2011). A review of 65 studies on the impact of ABF found no *consistent* systematic differences in readmission rates, mortality or volume of care between ABF and non-ABF systems. However, there were indications of increased severity of illness (Palmer et al. 2014), which could imply either that under ABF, more severe patients are treated (selected), or that more secondary diagnoses are registered. A review of 12 Scandinavian studies on activity-based reimbursement and efficiency also reported mixed results, and there was a larger effect in Sweden than in Denmark or Norway (Jakobsen 2010).

However, in Norway, there was a positive effect after the 1997 reform (Biørn et al. 2003), but not a very clear effect on early trials of ABF from 1990 to 1992 (Magnussen and Solstad 1994). While the Nordic countries' health systems share many traits, the use of DRGs has differed in several countries (Street et al. 2007). In Norway, all public hospitals face the same funding scheme with a large share of ABF. National guidelines also apply in Denmark, but with a very small share of ABF. County councils in Sweden and hospital districts in Finland choose how to fund hospitals, and consequently, there is large variation. Some Swedish counties use ABF,

as in Norway, while others employ no DRG-based ABF. In Finland, DRGs are used mostly for billing purposes among the hospital owners (municipalities).

Street et al. (2007) claim that the most important attractions to ABF are fairness and transparency. However, the funding set-up may impact how hospitals act by incentivising: reduced costs per patient, increased revenue per patient and increasing the number of patients (Cots et al. 2011). Incentives may also cause actors to form strategies to optimize outcomes under the funding scheme, as Neby et al. (2015) list a large array of DRG gaming strategies, including increasing the number of cases treated while decreasing their quality, selecting, dumping, creaming, skimping, skimming, undertreatment of patients, a revolving door effect, still-bleeding patient discharges, bribes, upcoding, overcoding, and case-splitting.

Pongpirul and Robinson (2013) categorize hospital manipulations in three groups: corporate, clinical and coding practices. They label all activity unrelated to patient care as corporate practices. Clinical manipulations then comprise three types: increasing admission volume, changing the intensity of care and the exaggeration of patient clinical conditions. The last group (coding) consists of many different types of behaviours: upcoding, overcoding, miscoding, coding optimization, coding practice change and code manipulation (Pongpirul and Robinson 2013).

The problem with upcoding as it relates to productivity measurement is that it will seemingly increase output without any proportional increase in input; thus, it will artificially raise productivity (Woolhandler et al. 2012). It has, however, also been argued that not all upcoding is fraudulent, but rather reflects an improvement in the quality of coding (Fisher et al. 1992; O'Reilly et al. 2012), indicating that past (pre-upcoding) estimates of productivity may have been underestimated if the output does not properly match the most resource-demanding patients. These studies show the possible effects of upcoding, and we will now consider some of the recent empirical papers on upcoding.

To start, Preyra analysed the coding responses to the introduction of a PPS in Canada and found increases in case-mix without any corresponding increase in resource use (Preyra 2004). Nonetheless, the results and consequences of upcoding differ depending on the setting. In the U.S., L. S. Dafny (2005) found a positive association between price incentives and upcoding. Increased severity of illness was found in a majority of studies on ABF, not just in the U.S. (Palmer et al. 2014). In Sweden, hospitals with a PPS experienced a larger increase in the

documentation of secondary diagnosis than hospitals with a block grant (Serdén et al. 2003). For-profit hospitals in the U.S. have also been found to be more receptive to upcoding (L. Dafny and Dranove 2009; Silverman and Skinner 2004), but this was not the case in Italy (Berta et al. 2010). In the public care setting of Portugal, Barros and Braun (2016) identified a positive association between price incentives and upcoding.

Liang (2015) recently presented the first study on upcoding outside of the U.S. and Europe. With data from Taiwan, she tested nine orthopaedic surgical DRGs and found a price effect, but also a cross-price effect (how the price of one DRG affects other DRGs). However, in Taiwan, not-for-profit hospitals were more responsive to profitability than for-profit hospitals (Liang 2015). Bowblis and Brunt found upcoding to be present in skilled nursing facilities funded by a PPS. However, given the possibility of audit, it was not patient severity that was upcoded, but rather the number of therapy minutes (Bowblis and Brunt 2014). A review showed that for-profit hospital ownership combined with the use of secondary diagnoses for reimbursement increased the potential for upcoding (Steinbusch et al. 2007). Likewise, auditing systems and the perceived risk thereof also affect the result and consequences of upcoding (Kuhn and Siciliani 2008).

4.6.1 Effects of the provider payment, research from Norway

Section 2.2 above presented details on hospital funding in Norway, which since 1997, has been in the form of a variable PPS (Jegers et al. 2002). The main argument for ABF reform was to reduce waiting times and increase hospital efficiency (Stortingsforhandling 1996; Street et al. 2007). However, not all applied the same logic *within* their counties, and although introduced in 1997, only 15 of the 19 counties immediately adopted the reform policy. The final counties implemented ABF only in 2000. Kjerstad divided hospitals into two groups (based on the time of implementation) and compared their development, finding that ABF "…gives stronger incentives to increase production compared to a block grant system" (Kjerstad 2003).

Elsewhere, Biørn et al. (2003) tested how ABF reform impacted hospital efficiency, and found improvements in technical efficiency, but less so for cost efficiency. In their sensitivity analysis, they controlled for a number of secondary diagnoses, which reduced the effect of the outcome measure, interpreting the difference as upcoding. Petersen and Anthun (2008) separated

different types of registration change in Norwegian hospitals over the period 2002–2005 (DRG logic, uncomplicated/complicated groups, unspecific main diagnoses, readmission and transfers), and found that the overall effect of registration change became larger each year compared with the demographically-caused changes in the case-mix. However, the use of secondary diagnoses had no impact on the case-mix index at the hospital level (Petersen and Anthun 2008).

In other work, Martinussen and Hagen (2009) examined cream skimming within DRGs by testing if waiting times were shorter for patients treated the same day instead of as inpatients. There was some evidence of cream skimming immediately after the introduction of ABF in 1997, but this did not increase further following the 2002 organizational reform. Tjerbo and Hagen looked at how the responses to the funding system not only depended on hospital behaviour at the patient level, but also on the political system. The setup of the Norwegian system resulted in a situation where the central state government sent signals of soft budget constraints and bailouts, and this led to increasing deficits (Tjerbo and Hagen 2009). In the years following their analyses, budget constraints have been set harder, and deficits have correspondingly decreased.

Biørn et al. (2010) tested whether hospitals responded homogeneously to the 1997 funding reform. They found that the response to funding reform was unrelated to the efficiency of the hospitals prior to the reform. Further, they suggested that budget constraints might be an explanatory factor (Biørn et al. 2010). Yin et al. (2013) tested how the level of ABF influenced length of stay for ischemic heart diseases using individual-level analysis of 331,046 hospital episodes. They found that a 10 per cent increase in ABF reduced the length of stay by 1.28 per cent, and that incremental changes in the ABF share were not enough to incentivise a reduction in costs or length of stay for that specific patient group.

Neby et al. (2015) presented several cases of DRG gaming in Norway and Germany. One of the Norwegian cases is a telling example of how upcoding might take place. A "…clinic had registered around 50 per cent of all patients as having undergone or being in need of tonsillectomies as snoring surgeries. A physician acting as an external consultant proposed a new 'creative' coding practice, adding a secondary diagnosis. He asked for a commission – 10 per cent of the extra funding" (Neby et al. 2015). The authors argued that even though they could not "…directly attribute gaming as such to NPM-style reforms, the role of political-administrative trajectories is clearly evident in how accountability practices play out" (Neby et

al. 2015). There have also been some studies on the accuracy and correctness of diagnostic coding (Burns et al. 2012; Campbell et al. 2001; Hsia et al. 1988). One Norwegian study found that as much as 37 per cent of all primary diagnoses were incorrect and that the number of secondary diagnoses was exaggerated (Jørgenvåg and Hope 2005).

Recently, Januleviciute et al. (2016) tested how hospitals responded to changes in the price gap in DRG pairs. They used data from a period of 5 years (2003–2007) and found an effect for medical patients, but not for surgical patients. Overall, a "…10 per cent increase in the ratio of prices between patients with and without complications increases the proportion of patients coded with complications by 0.3–0.4 percentage points" (Januleviciute et al. 2016). Melberg et al. (2016) looked at an issue related to upcoding when testing if changes in DRG weight were associated with increases in activity levels. The activity growth was higher in DRGs with price increases than in DRGs with price decreases.

While there are many studies on the Norwegian ABF system, there is a need for a greater understanding of the separate effects of ABF share, prices, and price changes. Most studies only aim to test one of these issues. Our study will thus add to the literature on hospital responses to prices using additional empirical tests and estimating the separate effects.

5 Aim

The aim of this thesis was to determine the development in hospital productivity and its relation to ABF in the hospital sector in Norway over the period 1999 to 2014. Specifically, we address four main issues: a) productivity and productivity changes during the period 1999 to 2014, b) a comparative analysis of productivity growth in the Nordic countries, c) the relationship between quality and productivity, and d) the relationship between hospital financing and diagnostic coding.

Paper I comprised an analysis of overall productivity growth in the Norwegian hospital sector over the period 1999 to 2014. We ensured comparability in the data by utilizing fixed weights and fixed grouper logic. Paper I aimed to estimate the total productivity development during the period and to decompose the change into frontier shifts and technical efficiency change. By also estimating scale efficiency, the paper aimed to compare the estimated optimal scale as opposed to the actual hospital size.

Paper II compared hospital productivity across four Nordic countries. Earlier comparative studies indicated that Finnish hospitals had higher productivity than other Nordic countries (Linna et al. 2006; Kittelsen et al. 2007; Kittelsen et al. 2008; Kittelsen et al. 2009; Linna et al. 2010; Kalseth et al. 2011). Paper II aimed to decompose productivity in the period 2005–2007 to determine if there were country-specific frontiers of efficiency that could explain country differences, or if any differences were largely because of differences behind the frontier.

Paper III represented a continuation of Paper II by examining if quality differences (as measured by quality indicators) between the main Nordic countries could explain the observed differences in productivity. Paper III also aimed to estimate if there were any trade-offs between quality indicators and productivity. To do this, we linked data for the years 2008 and 2009 to out-of-hospital mortality to not only examine in-hospital outcomes, but also follow up patients after treatment.

Productivity growth could be also a consequence of changes in the coding of diagnoses and procedures. Paper IV aimed to analyse if there was an association between the level of upcoding and the potential economic gain within DRG pairs during the period 1999 to 2008.

6 Methods and materials

6.1 Data

The primary source of data for all four papers was the Norwegian Patient Registry (https://helsedirektoratet.no/english/norwegian-patient-registry). The registry contains complete records of all specialized hospital episodes in Norway, including all inpatient admissions, day care and outpatient treatments in Norwegian acute care hospitals over the period 1999 to 2014. We used a subsection of the data in each of the four papers. The data included treatments in both public and private hospitals with long-term contracts with RHAs. For each hospital episode, the data contained information about the patient (age, sex), the episode (dates and times of admission and discharge), the hospital (hospital trust, hospital and department), the treatment (procedures and diagnoses) and details related to the ABF (DRG, type of DRG and DRG weight).

Each hospital episode was assigned to a DRG (by the Norwegian Patient Registry) based on the diagnoses and procedures recorded, as well as patient age and sex and length of stay. We used length of stay to determine if a hospital visit was an outpatient consultation, day care or an inpatient admission. There are currently approximately 780 different DRGs. After each episode is DRG grouped, it is also assigned to a cost weight (also called a DRG weight), which is a relative measure of the average cost for treatment in each DRG. Using these weights, we aggregated the activity data into composite output measures at the hospital level. For two of the papers, we regrouped (see Section 6.1.1, Section 6.1.3 and Appendix (Chapter 10) below) the data to facilitate both longitudinal and international comparisons. Regrouping with a common grouper removes year or country-specific grouper effects.

Of course, it would have been beneficial to include data preceding the introduction of ABF in 1997. However, in 1999, the classification of diseases changed from version ICD9 to ICD10 (WorldHealthOrganization and StatensHelsetilsyn 1998). Determining comparable output measures from before 1999 is thus very difficult, and it was unfortunately not possible within the scope of this thesis to extend the analysis further than a 16-year span.

6.1.1 Paper I

Paper I used patient administrative data from 1999 to 2014 (see previous section). All radiotherapy treatments (DRG 409O) were excluded, as these treatments were not part of the costs. Healthy new-borns (DRG 391) were not registered in 1999–2001, but generated in the data using public statistics on the number of births.

Each year, the rules for DRG grouping have changed. To allow for a longitudinal comparison of output data, the data were regrouped using a common DRG logic definition for the whole sample period. A DRG grouper is software that assigns each patient's diagnoses and procedures into a DRG. The software uses patient data as inputs. Based on patient length of stay, sex, age, discharge status, diagnoses and procedures, each episode was assigned into a DRG. The Norwegian Patient Register routinely does this DRG grouping, but only with annual grouper versions; thus, for instance, episodes in 1999 will be grouped with the 1999 grouper logic. As the grouper logic changes over time, having one common version would remove any grouper logic effect from the measurement of hospital output. In short-term comparisons, this issue may be negligible, but in the long run, this is important to control for, and thus, Paper I incorporated a 16-year time span.

Cost weights were calculated as mean weights per DRG in 2011. Some rare groups did not have any treatment in 2011 (and thus no cost weights), and for these, actual weights were used. As a sensitivity test (results not shown), we also estimated results using fixed 2014 and annual weights, but fixed weights give the best results in terms of allowing technical change, and as the grouper is 2011 based, we chose 2011 weights. After regrouping, all episodes were aggregated into four composite types of hospital activity: 1) emergency inpatient discharges, 2) elective inpatient discharges, 3) day care treatments and 4) outpatient visits and treatments. These four measures served as our output dimensions in the productivity analyses.

Ideally, inputs should be measured as actual inputs used in production such as personnel, beds and medication. However, these were not available from comparable sources for similar definitions for the entire period of our study, and according to Jacobs et al. (2006):33, "If a longer-term, less constrained analysis is required, then a single measure of 'total costs' may be a perfectly adequate indicator of physical organisational inputs". Hospital inputs were measured

as total operating costs. These were defined identically to costs presented in the national annual reports on hospital production in Norway (Samdata 2015), and the sources of the data were Statistics Norway and the Norwegian Directorate of Health. The costs were adjusted to include only patient-related costs. Capital costs were deducted, as these were only comparable over the period 2004–2014. Costs for teaching and research were removed, as well as other incomes not relevant to DRG production.

To allow for comparison over time, the costs had to be deflated to fixed prices. This was done using the same deflator as the public SAMDATA reports (Samdata 2015) for the years 2005-2014. For the period 1999–2004, the deflator formulated by Kittelsen and colleagues (Kittelsen et al. 2007). Table 2 details the annual price increases. The effect is cumulative so that the value of 1 NOK in 1999 is increased 104 per cent to be comparable with 1 NOK in 2014.

Table 2 Annual deflator, increase from previous year

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
7.0	2.9	8.1	6.9	5.2	2.9	4.2	8.0	6.0	2.7	2.5	5.5	4.0	3.9	3.5

6.1.2 Paper II

Paper II employed data from 2005 to 2007. In addition to the data specified in Section 6.1.1 above, Paper II also used comparable data from the other major Nordic countries collected by participants in the Nordic Hospital Comparison Study Group (Medin et al. (2013) and http://www.thl.fi/en_US/web/en/research/projects/nhcsg). Swedish data were only available for 2005 and 2006, and hospital costs in Sweden were only available at the administrative county level. Consequently, we aggregated the Swedish activity data to this level for the analysis.

Only acute care hospitals were included; private for-profit hospitals were excluded, as well as all Finnish health centres, which in some cases had inpatient treatment, but were otherwise not comparable to hospitals in the other major Nordic countries. Some patient groups had to be excluded, as they were not considered hospital treatment in all countries (healthy new-borns, rehabilitation, dialysis, radiation therapy and chemotherapy). Further, there are differences

between countries in how diagnostic coding is done and how much ABF is used in budgeting. Some DRGs were linked in pairs of complicated and uncomplicated treatment/patients of the same diagnoses. The addition of one or more secondary diagnoses would cause the patient to be grouped in complicated instead of uncomplicated groups. To reduce the impact of national differences in coding practices, we aggregated the DRG pairs from multiple to single DRGs. This necessarily reduces the output of hospitals with more secondary diagnoses than average. We defined three output categories: 1) inpatient care, 2) day care and 3) outpatient visits. Within each category, the treatments and patients were weighted with Norwegian cost weights (from 2007). The cost weights reflected the average cost of treatment for each group. The Danish DRG system was not directly comparable to the other Nordic countries, and Danish DRG weights were used for the specific Danish DRGs at the same time as the level was normalized to the mean Nordic cost weights.

Hospital costs were the only input, and were similarly defined in all four countries to include only production-related costs (Anthun et al. 2013). Capital costs were not used, as there were substantial differences in the accounting rules on how these were depreciated in each country. In addition, personnel statistics, which might be a good measure of hospital inputs, were not used, as these data were not available from all countries. To harmonize the cost level between the countries, cost indices were constructed (Anthun et al. 2013; Kittelsen et al. 2009). These indices were based on wages for physicians, nurses (and other groups of hospital personnel) and a purchaser parity-corrected GDP price index from the OECD.

Furthermore, some hospital-level variables were collected to account for environmental factors and to control for some case-mix indices perhaps not captured by the DRGs (see Table 3).

Table 3 Hospital-level explanatory variables for the analysis in Paper II

Variable	Definition	Reasoning and expectation
University hospital	Dummy for university hospital status	Costs for teaching and research are excluded (see description in Section 6.1.1), but there might be related scope effects. Also, university hospitals may possibly treat more severe patients
Capital city hospital	Dummy for hospitals located in capital city	Different socio-economic composition in catchment area, different travelling times and greater potential for outpatient treatments
Case-mix index	Average DRG weight per patient	If case-mix differences are not fully captured by the DRG system, we would expect lower productivity in hospitals with more severe and treatment-intense case loads
Length of stay deviation	DRG weighted average LOS in each DRG for each hospital divided by the average LOS in each DRG for all hospitals	Could capture severity not captured by the DRGs and case-mix index, but may also capture inefficiency
Outpatient share	Share of output that is outpatient	We assumed that a high output share may indicate lower costs per discharge

6.1.3 Paper III

Paper III draws on hospital activity and cost data from Nordic countries over the period 2008–2009. The paper also includes several quality indicators. These data were collected by participants in the Nordic Hospital Comparison Study Group (http://www.thl.fi/en_US/web/en/research/projects/nhcsg and Medin et al. (2013)). Patient level data from all countries were collected in each country before an anonymous subset of the data was submitted centrally to a secure server (at the Ragnar Frisch Centre for Economic Research in Oslo), where the joint and pooled analyses of the four countries were performed based on 58,158,847 records.

We used readmissions, out-of-hospital mortality and patient safety indicators as defined by the OECD (Drösler 2008) as comparable quality indicators throughout the Nordic countries. Table 4 below lists and defines these variables. Cost and patient administrative data were defined identically as in Paper II; however, in Paper III, the data were regrouped using a grouper with fixed definitions. This was done for all countries to ensure comparability, especially Denmark, which in Paper II was grouped in the local Danish DRG version (DkDRG), which is not directly comparable to the NordDRG version used in Norway, Sweden or Finland. To aggregate data,

new cost weights were calculated from pooled 2008 and 2009 cost per patient information from two Finnish hospital districts and the regrouped Finnish activity data.

Table 4 Quality indicators used in Paper III

Variable	Definition
Acute readmission	Patient admitted acutely to inpatient care within 30 days of discharge
Inpatient readmission	Patient admitted to inpatient care within 30 days of discharge
Out-of-hospital 30-day mortality	Dummy for last hospital episode for deceased patients within 30 days of admission
Out-of-hospital 90-day mortality	Dummy for last hospital episode for deceased patients within 90 days of admission
Out-of-hospital 180-day mortality	Dummy for last hospital episode for deceased patients within 180 days of admission
Out-of-hospital 365-day mortality	Dummy for last hospital episode for deceased patients within 365 days of admission
Patient safety indicator: Pulmonary embolism or deep vein thrombosis	Dummy for presence of secondary diagnosis: I26.0, I26.9, I80.1, I80.2, I80.3, I80.9, I80.9, I82.8, I82.9
Patient safety indicator: Sepsis	Dummy for presence of secondary diagnosis: A40.0, A40.1, A40.2, A40.3, A40.9, A41.0, A41.1, A41.2, A41.3, A41.4, A41.5, A41.8, A41.9, R57.8, T81.1
Patient safety indicator: Accidental cut, puncture, or haemorrhage during medical care	Dummy for presence of secondary diagnosis: T81.2, Y60.0, Y60.1, Y60.2, Y60.3, Y60.4, Y60.5, Y60.6, Y60.7, Y60.8, Y60.9
Patient safety indicator: Obstetric trauma	Dummy for presence of secondary diagnosis: O70.2, O70.3
Patient safety indicator: Bed sores	Dummy for presence of secondary diagnosis: L89

Information on out-of-hospital mortality for Norwegian patients was collected from the National Register and linked by the Norwegian Patient Register. All quality indicators are based on the same patient administrative data that form the basis of the output data. Readmission variables were calculated based on admittance and discharge dates. Out-of-hospital mortality was measured as the distance (in days) to death (only for those who die). Patient safety indicators are dummy variables defined by the coding of certain specific secondary diagnoses.

The paper also utilizes further case-mix adjustment variables at the individual, municipal (based on the home municipality of the patient) and hospital levels, as listed in Table 5.

Table 5 Explanatory variables in Paper III

Type	Variable	Definition
Individual/patient/episode	Sex	
	Age in groups (years)	0, 1–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90+
	Transfers	Four dummy variables created for transfers into and out-of-hospital department stay within one day before or one day after each episode
	Charlson	Charlson index based on secondary diagnoses (Charlson et al. 1987)
	Number of secondary diagnoses	Number of secondary diagnoses
	Length of stay	Defined as discharge date – admission date + 1
Municipality	Population	Population of patient home municipality
	Unemployment rate	Unemployment rate as % of labour force
	Social assistance	Social assistance recipients as % of population
	Single-parent families	Single parent families as % of all families with children
	Foreign country	Citizens of foreign countries as % of population
	Travelling time	Travelling time by car in hours between hospital and centre of home municipality
Hospital	Average cost per DRG point	Costs divided by total DRG points
	Number of patients	Number of departmental/speciality discharges
	Case-mix index	Hospital DRG points divided by number of patients
	University hospital	Dummy if hospital is a teaching or university hospital
	Capital city hospital	Dummy if hospital is located in the country's capital city

6.1.4 Paper IV

The purpose of Paper IV was to analyse the relationship between hospital financing and diagnostic coding. This was done by a test of whether the upcoding was associated with price incentives. Paper IV used data on inpatient discharges from 1999 to 2008; however, these data were not grouped to a common version grouper as in Papers I and III. A common grouper would create new groups in historical data and impose a non-existent incentive structure on the actors. Instead, we selected the data to include only groups that we assume are comparable. We selected a sample of DRG pairs where the addition of one or more diagnoses or procedures would have resulted in the patient being classified as complicated instead of uncomplicated, according to the specification of NordDRG and the Directorate of Health. Only DRG pairs that existed throughout the entire period and only those with more than 1,000 cases annually were included in this analysis. This resulted in 76 DRG pairs based on 3,180,578 inpatient discharges. The data were aggregated to the hospital-year–DRG pair level (N=19,250) to allow for a three-level mixed-method approach (see Section 6.2.2 below). The sample period of Paper IV was limited to 1999–2008 because of substantial changes in ABF. The other papers circumvented

this issue by regrouping the data. For Paper IV, regrouping was not an option because it could potentially regroup discharges into DRGs with different incentives for upcoding.

Table 6 below lists the variables used in Paper IV and the level of observation.

Table 6 Variables in Paper IV

Variable	Level of observation	Definition
Age	Hospital–year–DRG pair	Mean age
Sex	Hospital–year–DRG pair	Percentage female
Emergency	Hospital–year–DRG pair	Percentage of episodes as emergency
Length of stay	Hospital–year–DRG pair	Mean length of stay
Number of inpatient treatments at hospital	Hospital–year	Case-mix adjusted
Medical DRG	DRG pair	DRG type medical or surgical (dummy for medical)
Charlson co-morbidity index	Hospital–year–DRG pair	Charlson index (not based on those diagnoses that result in complicated status)
Potential gain in income	DRG pair	Main incentive variable. Calculated as mean price difference between complicated and uncomplicated group multiplied by annual ABF share
Changes in potential gain in income	Year–DRG pair	Changes in main incentive variable. Difference between yearly potential gain in income and mean potential gain in income per DRG pair
Percentage of complicated discharges	Hospital–year–DRG pair	Dependent variable, how large share of treatments in DRG pair are in complicated group
Time trend	Year	Time trend (in years) since 1999
Time dummy	Year	Dummy for the years post-reform 2002-2008
Trend and reform interaction	Year	Interaction time trend and reform-dummy

6.1.5 Summary of data and data sources

Table 7 below summarizes the type of data used in each paper.

Table 7 Type of data, years covered by study, source and variables, by paper

Type of data	Paper I	Paper II	Paper III	Paper IV	Source
Patient administrative data	1999–2014	2005–2007	2008–2009	1999–2008	Norwegian Patient Registry
Cost	1999–2014	2005–2007	2008–2009		Statistics Norway
Explanatory variables		Hospital, 2005–2007	Individual, hospital and municipal, 2008–2009	1999–2008	Norwegian Patient Registry, Statistics Norway
Quality indicators			2008–2010		Norwegian Patient Registry, National Register
Cost weights	Fixed 2011 weights	Fixed 2007 Norwegian weights	Fixed Finnish weights (constructed based on pooled 2008–2009 regrouped activity and cost-per-patient data)	Annual weights	Norwegian Patient Registry
Grouper version	Fixed grouper	Annual national grouper	Fixed common Nordic grouper	Annual grouper	

Note: Papers II and III have other sources for the other Nordic countries

6.2 Methods

6.2.1 Measuring efficiency and productivity

In Papers I, II and III, we undertook data envelopment analyses (DEA) with bootstrapping, and for Papers I and II, we also decomposed the DEA results using Malmquist indices. Paper II estimated the cost frontier and inefficiency, using not only DEA, but also SFA.

Data Envelopment Analyses (DEA)

A voluminous number of studies concern productivity and efficiency in health care systems and hospitals. These studies employ many different techniques and methods. O'Neill et al. (2008) presents two important methodological dimensions for hospital efficiency studies:

parametric/non-parametric and deterministic/stochastic methods. Parametric methods assume a specific functional form of the frontier. Deterministic methods assume that all deviation from the frontier is inefficiency, while stochastic methods produce random errors as part of the estimation and are thus less sensitive to outliers.

Inefficiency can be defined as "…the extent to which an organisation's…output falls short of that predicted by the production function…" and thus, it is inherently unobservable (Jacobs et al. 2006). Inspired by Farrell (1957), Charnes et al. (1978) created the linear programming now known as DEA, which offers a pragmatic solution to this problem. First, measure observable phenomena (inputs and outputs), and then use these to estimate a productivity frontier, being the best output for a certain level of input. Second, define each unit's efficiency as the ratio of its observed data and the productivity frontier. As we cannot know *a priori* what the best possible productivity is, the DEA approach defines best possible as best observed.

The DEA method is non-parametric, which implies that the observed data infers the shape of the frontier. Conversely, a parametric approach implies a certain form of the frontier and is most often associated with the estimation of cost functions. DEA has three major assumptions: 1) that all observed units are feasible, 2) free-disposability and 3) convexity (O'Neill et al. 2008). The primary weakness of the DEA method is that it assumes no measurement error. If there is only one output and one input, there is no need for advanced methodologies, as productivity is simply a ratio of the output to the inputs (Castelli et al. 2015; Jacobs et al. 2006). However, it is common in health care settings to have multiple inputs and outputs, and the method allows easily for multiple inputs and outputs by weighting their combinations. The weights specific for each observation are such that the input–output ratio is maximized relative to all other observations (Jacobs et al. 2006).

The estimation of DEA scores is done by solving for each unit in the following equation (Jacobs et al. (2006)):

$$\max\left(\frac{\sum_{s=1}^{S} u_s \times y_{s0}}{\sum_{m=1}^{M} v_m \times x_{m0}}\right) \tag{6}$$

where y_{s0} is the quantity of output s for hospital, x_{m0} is the quantity of input m for hospital$_0$, u_s is the weight attached to output s and v_m is the weight attached to input m. The equation is subject to the constraints that no hospital can have greater efficiency than one, and all weights must be positive.

Simar and Wilson (1998, 2000) proposed bootstrapping by repeatedly simulating the data-generating process. This allows us to obtain a sampling distribution of the estimates and to estimate the bias and CIs of the estimators. The number of bootstrap iterations in Papers I, II and III was 2,000, except for the scale efficiency estimators in Paper I, which were bootstrapped with 1,000 iterations (due to limited computational resources). All efficiency scores presented in Papers I, II and III were corrected for the bootstrap bias, and all CIs were 95 per cent or better. Bootstrap-corrected estimates have fewer observations on the frontier because there are likely fewer efficient units when bias is corrected.

When decomposing productivity in the Nordic countries (in Paper II), we also estimate the national frontiers to estimate if there are country-specific technologies. This is done by enveloping the time-specific technologies and treating this envelopment as a technology. By comparing each country's mean frontier with the overall frontier, we estimate each country's productivity.

Optimal scale

DEA can be used to calculate the optimal scale, as developed by Banker (1984) and Banker et al. (1984). Paper II calculated the scale elasticities based on DEA. Scale efficiency is a measure of the distance from the VRS to the CRS frontier, given the observed levels of inputs and outputs. We estimate optimal hospital size by calculating the scale properties of hospitals compared to those hospitals that are scale efficient. This was achieved in Paper I by performing annual DEA analyses based on accumulated data. The optimal size of each hospital was estimated by multiplying the observed productivity of each hospital by its cost and then dividing by the relative distance to the optimal scale-efficient hospital. A CI is estimated based on the distribution of the bootstrapped estimate of the relative distance to the optimal scale-efficient hospital (Simar and Wilson 1998, 2000).

Measuring change

The Malmquist index was proposed by Caves and colleagues (Caves et al. 1982) as a way to estimate the productivity differences between firms by substituting the technology (inputs and costs) they face. Färe and colleagues demonstrated how it was possible to decompose the Malmquist index to show both the frontier change and the efficiency change (Färe et al. 1992), and in doing so, they bypassed "…the need for price information and allowing for inefficiency, while preserving the requirement that the framework holds for very general production structures" (Färe et al. 1994a). In Papers I and II, the Malmquist indexes of productivity growth were estimated by comparing the DEA results for the same DMU over two years. For the Malmquist technical change component, we assumed sequential (accumulated) frontiers. The productivity for a unit may be compared with best-practice hospitals at the frontier in the same and earlier years, but never with observations in the future.

Under a CRS assumption, the Malmquist index is a product of technical change (frontier shifts) and efficiency change. Given that CRS technical efficiency is the product of VRS technical efficiency and scale efficiency, Malmquist decomposition can easily be extended to estimate the scale efficiency change (Färe et al. 1994b; Lovell 2003), as in Paper II.

Stochastic Frontier Analysis (SFA)

Two of the most common estimation methods used for hospital productivity analyses have been DEA and SFA. However, by 2008, less than one-fifth of the studies reviewed used SFA (Hollingsworth 2008). Comparisons have been made between the methods by a few select studies, and the results are reported to correspond to some degree (Jacobs 2001; Linna and Häkkinen 1999; Mortimer 2002). Jacobs (2001) concluded "…each of the methods does have unique strengths and weaknesses and potentially measures slightly different aspects of efficiency". SFA is basically a regression method estimating a cost function where (in)efficiency is measured as part of the error term (Jacobs 2001).

Paper II calculates cost frontier and inefficiency using not only DEA, but also SFA. SFA is based on the method outlined by Battese and Coelli (1995), and estimates cost functions using maximum-likelihood estimation. Because SFA uses normal regression models, it is possible to test different technological assumptions, functional forms, and distributions of the error terms

easily. Such tests were performed in Paper II to select the properties that gave the best regression fit.

Association between quality indicators and productivity estimates

Paper III aimed at developing patient register-based measures of quality, and accomplished this by estimating to what extent case-mix adjusted quality indicators were associated with the productivity estimates for hospital trusts. This was attempted by calculating the observed-to-expected ratio of each quality indicator for each hospital. Each patient had a (binary) observable quality indicator (and an expected quality indicator). The case-mix adjusted hospital performance measures were calculated by summing all observed patient outcomes and dividing by the sum of all expected patient outcomes.

Different models perform case-mix adjustment differently; in its simplest form, it is merely the average value of the quality indicator within each DRG for all patients across all hospitals. We estimated the expected value based on maximum-likelihood estimation of the quality measure for patient i in DRG k at hospital h. The individual predicted values were used to calculate a hospital-specific index for each quality indicator. Hospital-level pairwise Pearson correlation coefficients for each performance indicator, average costs (per DRG point) and productivity estimates were estimated, along with a simple regression of the DEA bootstrapped estimate of productivity for hospital h in country c in year t, including the performance indicators as explanatory variables to estimate the quality–productivity trade-off.

6.2.2 Regressions

To perform statistical modelling and testing, each paper employed different forms of regression. The different regression models were based on the structure of the data and the nature of the research question at hand. These models were as follows:

- Ordinary least square second-stage regressions to test which variables were associated with the productivity estimates in Paper II

- Maximum-likelihood estimation of the cost functions in Paper II to estimate the inefficiencies in SFA
- Generalized least squares random hospital effects in Paper III to estimate the productivity–quality trade-off
- Multi-level linear regressions in Paper IV to test the impact of prices on hierarchical data across three levels

6.2.3 Software

All DEA analyses (in Papers I, II and III) were executed using the *FrischDEA* package provided by the Ragnar Frisch Centre for Economic Analysis. All Malmquist analyses in Papers I and II were executed using the *FrischMalmquist* package provided by the Ragnar Frisch Centre for Economic Analysis.

The patient-level data utilized in Papers I and III were regrouped to a fixed NordDrg grouper version using a *DRG grouper (*version *NOR2011co1F* in Paper I and version *DRGL_2008_MV Nor 09/02/20* in Paper III*)* provided by DataWell Oy.

All statistical analyses were performed using *Stata*.

6.2.4 Summary of methods

Table 8 below summarizes the methods applied in each paper.

Table 8 Summary of methods applied in Papers I–IV

Paper	DEA bootstrapped	Malmquist decomposition	SFA	Regressions
Paper I	X	X		
Paper II	X	X	X	Second-stage cost functions
Paper III	X			Quality trade-off
Paper IV				Multi-level

6.3 Ethical considerations

The project was approved by the Regional Ethical Committee for Research (Approvals 2009/999, 2009/1610, 2011/930B and 2012/1887), the Norwegian Centre for Research Data (acting as Data Protection Official for Research), the Norwegian Data Protection Authority, and the Norwegian Directorate of Health. These approvals allowed the collection of data on hospital activity without the consent of patients. Even though the data include details of individual-level treatments, it is impossible to identify individual patients in the data without significant prior knowledge of the patient's treatment. All analysis has been on either aggregate data or anonymized subsets of data.

7 Summary of results

7.1 Summary of Paper I

Paper I aimed to estimate the total productivity development in Norwegian hospitals over a period of 16 years, 1999 to 2014, and to decompose the development into frontier shifts and technical efficiency. By also estimating scale efficiency, the paper aimed to compare actual with estimated optimal hospital size. Hospital admissions were grouped into DRGs using fixed grouper logic. Four composite outputs were defined, and inputs were measured as operating costs. Productivity and efficiency were estimated using bootstrapped DEA.

Mean productivity increased by 24.6 percentage points from 1999 to 2014, an average annual change of 1.5 per cent. There was substantial growth in productivity and hospital size following ownership reform. After the reform (2003–2014), average annual growth was <0.5 per cent. There was no conclusive evidence of technical change. Estimated optimal size was smaller than the actual size of most hospitals, yet scale efficiency was high, even after hospital mergers.

7.2 Summary of Paper II

Paper II aimed to compare and decompose the hospital productivity between the major Nordic countries. The purpose of the paper was to determine if there were country-specific frontiers for efficiency. Previous studies indicated that Finnish hospitals had significantly higher productivity compared with other Nordic countries (Kittelsen et al. 2008; Linna et al. 2006; Linna et al. 2010). As there was no natural pairing of observations between countries, we estimated productivity levels, rather than a Malmquist index of productivity differences, using a pooled set of all observations as a reference. We decomposed the productivity levels into technical efficiency, scale efficiency and country-specific possibility sets (technical frontiers). Data were collected on operating costs and patient discharges in each DRG for all hospitals in

the four major Nordic countries, Denmark, Finland, Norway and Sweden for the period 2005–2007.

The mean productivity in Norway, as measured against a common pooled reference frontier, was 56.6 per cent. This indicates that compared with the best Nordic hospital(s) in the three years of the study, the Norwegian hospitals were 43.4 per cent behind the frontier on average. When decomposing this productivity, the study found that scale efficiency in Norway was higher than that in Finland, with an average scale efficiency of 94.2 per cent. However, the scale elasticity was less than one, indicating that Norwegian hospitals on average were too large. The technical efficiency, i.e., the mean efficiency behind the Norwegian frontier, was also high, at 89.7 per cent, indicating that the country-specific frontier resulted in low overall productivity.

The common frontier was fully composed of Finnish hospitals, and the country-specific Finnish frontier was the main source of the Finnish productivity advantage. The Norwegian frontier was more than 30 per cent behind the Finnish frontier. There were small differences in scale and technical efficiency between the countries, but large differences in production possibilities (the frontier position). There was no statistically significant association between efficiency and hospital status as a university or capital city hospital. The results were robust to the choice of either bootstrapped DEA or SFA as the frontier estimation methodology.

7.3 Summary of Paper III

The paper developed and analysed patient register-based measures of quality for the major Nordic countries. Previous studies (including Paper II) showed that Finnish hospitals had substantially higher average productivity than hospitals in Sweden, Denmark, and Norway, and that there was substantial variation within each country (Kittelsen et al. 2008; Linna et al. 2006; Linna et al. 2010). Paper III examined if quality differences between the major Nordic countries could explain the differences in productivity, and Paper III also attempted to test if there were any trade-offs between productivity and quality indicators.

Data on costs and discharges were collected for 160 acute care Nordic hospitals over the period 2008–2009. The patient register-based measures of quality were readmissions, mortality and patient safety indices. Patient safety indices were created based on observed and expected

incidents at the patient level, and case-mix adjusted at the DRG level. Productivity was estimated using bootstrapped DEA.

There were significant differences in the case-mix adjusted performance measures, as well as in productivity at both the national and hospital levels. For most quality indicators, the performance measures revealed room for improvement. The ranking of the country means of the case-mix adjusted performance measures varied across the different measures. Norway had higher readmission rates than the other countries, but the lowest mortality rates. Norwegian hospitals had the lowest rate (i.e., higher performance) for two of the patient safety indicators: PSI12 (pulmonary/deep vein thrombosis) and PSI 13 (sepsis); however, Norway also had the lowest performance for PSI 15 (accidental cut, puncture, or haemorrhage during medical care).

This paper confirmed the relative ranking of Nordic hospitals from earlier studies, with Finnish hospitals being more productive on average; however, Denmark was almost as productive. Norwegian hospitals seem to have caught up somewhat compared with Paper II. There was a weak but statistical significant trade-off between productivity and inpatient readmissions within 30 days, but hospitals with high 30-day mortality also tended to have higher costs. Hence, no clear cost–quality trade-off pattern was revealed.

7.4 Summary of Paper IV

Estimated productivity growth could be a consequence of altered coding of diagnoses and procedures rather than altered practices. Paper IV aimed to analyse if upcoding quantifiable at the national level was associated with the implicit price incentive for upcoding within DRG pairs.

The funding of day care and outpatient care was reformed in 2009, so Paper IV examined the years 1999–2008 by analysing 3,180,578 hospital discharges. We examined pairs of DRGs where the addition of one or more specific diagnoses placed the patient in a complicated rather than in an uncomplicated group, yielding higher reimbursement. The economic incentive was measured as the potential gain in income by coding a patient as complicated, and we analysed the association between this gain and the share of complicated discharges within the DRG pairs.

Using multi-level linear regression modelling, we estimated both differences between hospitals for each DRG pair and changes within hospitals for each DRG pair over time. Over the entire period, a one-DRG point difference in price was associated with an increased share of complicated discharges of 14.2 (95 per cent CI, 11.2 to 17.2) percentage points. However, a one-DRG point change in prices between years was only associated with a 0.4 (95 per cent CI, –1.1 to 1.8) percentage point change of discharges into the most complicated diagnostic category. Although there was a strong increase in complicated discharges over time, this was not closely related to price changes as expected. However, when stratifying on medical/surgical DRG pairs, there was a positive association between the change in prices and the share of complicated discharges for medical DRG pairs, and a negative association for surgical DRG pairs.

8 Discussion

8.1 Discussion of the results

Biørn et al. (2003) show that hospitals, on average, improved their technical efficiency with the introduction of ABF, and this corresponds to arguments that DRG-based hospital payments enhance efficiency (Street et al. 2011). However, what results follow after the effects from the introduction of ABF? In Paper I, we find overall productivity growth of 24.6 per cent from 1999 to 2014, an average annual improvement of 1.5 per cent. Given that the data for 1999–2014 determines the frontier and subsequently the mean level of efficiency behind the frontier, direct comparisons of productivity *level* across studies is not feasible. In addition, earlier (Norwegian) studies on hospital productivity development cover fewer years (Biørn et al. 2003; Martinussen and Midttun 2004; Hagen et al. 2006; Kittelsen et al. 2008), but the relative development in our study seems to be comparable. Our study is, to our knowledge, the first hospital-level productivity study covering such a long period.

There is also a possibility that the way the ABF system is set up might affect productivity. As reimbursements are based on national average costs (DRG weights), it follows that the local cost level hospitals face may affect their measured productivity. Given we have only used costs as inputs, hospitals with low average costs will seem productive, even if they employ the same amount of personnel as other hospitals. Previous analysis has shown that there are great differences in the cost level in Norway. A Norwegian Official Report (NOU 2008:2) found that the level of costs at the costliest hospital was ~60 per cent higher than the least costly hospital. The redistribution of income between RHAs takes some of this into account. Our analysis in Paper I did not control for differences in costs or prices.

Caves et al. 1982 wrote in their seminal Malmquist paper: “Thus the empirical usefulness of the Malmquist indexes is limited”, only 10 years later to be proven wrong by Färe et al. (1992), who used Malmquist indexes to decompose frontier and efficiency change; bypassing information about prices further increased their empirical usefulness. That being said, in our first paper, the results from the decomposition do not tell an interesting tale of the development

as no discernible pattern emerged: two years involved significant frontier shifts, four years catching up, and two years falling behind. The Malmquist index of productivity growth yielded somewhat clearer results for six years of growth and three years of decline.

While numbers are growing, there is a lack of good and comprehensive comparative studies. Papers II and III add to the literature by providing studies with a thorough comparative perspective, but this would not have been possible if the countries did not have such similar health systems to start with. As we recall, Dervaux et al. (2004) found it impossible to compare the U.S. and France with the same technology. Papers II and III show that such comparison is possible, but that there still seems to be country-specific frontiers, as determined in Paper II. In addition, the implications mentioned in Paper III of treatment patterns and practices varying significantly across countries corresponds to this observation. We did not identify a clear pattern that any country has higher or lower quality across all quality measures.

Our findings confirm those in other studies (Kalseth et al. 2011; Kittelsen et al. 2009; Kittelsen et al. 2008; Linna et al. 2006; Linna et al. 2010; Medin et al. 2011; Medin et al. 2013) that Finnish hospitals are more productive than hospitals in the other major Nordic countries. All the Finnish advantages appear to stem from the country-specific frontier, and similarly, all the Norwegian disadvantages owe to the Norwegian frontier because scale and technical efficiency is just as high in Norway as in the other countries. In line with earlier studies, the Swedish hospitals were the least efficient. What institutional and policy settings account for the country-specific frontier we can only speculate, but those issues, including quality, have not been tested thus far.

8.1.1 Optimal size

Street et al. (2007) assert, "[p]ublic hospitals operate in constrained environments. They cannot choose where they are located, or the population they serve, and…they have limited discretion about their size and the mix of specialities they have". These issues will obviously affect the cost level each hospital faces, regardless of its productivity level. Norway is very sparsely populated, which only exaggerates this issue. With a political consensus for maintaining population levels in rural areas, it is a balancing act to supply large hospitals and maintain

smaller hospitals or to have good ambulance services. It is not obvious *a priori* that all Norwegian hospital services enjoy optimal scale.

We have followed the advice of Dranove and Kristensen et al. (Dranove 1998; T. Kristensen et al. 2012) and not estimated the optimal number of beds. As hospitals rely less and less on inpatient treatment, the number of beds is perhaps an outdated measure of hospital size. We estimate (in Paper I) the optimal size of each hospital, as some hospitals are too small and others are too large. Our results showed an optimal size of 223 million NOK (in real 2014 NOK) in 1999, and this shifted up dramatically following hospital reform to 496 million NOK in 2003; in 2014, we estimate this to be 629 million NOK. Unfortunately, the CIs are quite large, so the precision of these estimates is low. However, the most interesting fact is not the estimated optimal size, but rather the fact that the estimate is very low compared with the actual size of Norwegian hospitals. By 2014, no hospitals were smaller than the estimated optimal size. Ideally, the estimated size would be the median observation. It must be noted that after hospital reform, the data are at the hospital trust level, which are large aggregates as very often they multi-hospital structures.

Førsund and Hjalmarsson (2004) conclude that under DEA, optimal scale can be found at various points in the input/output space. While we find optimal scale to be quite small, it must be noted that Norwegian scale efficiency is high, as shown in both Papers I and II. In the period 1999–2014, scale efficiency was an average (arithmetic mean across years) of 95.4 per cent, and despite a low optimal size, the scale efficiency was marginally higher in 2008–2014 than in 1999–2007. When measured against the multi-country Nordic frontier in 2005–2007, the mean scale efficiency in Norway was 94.2 per cent, which was higher than the Finnish or Danish scale efficiencies.

Preyra and Pink assert that their model was not statistically as good for the largest hospitals (Preyra and Pink 2006), and Paper II found diseconomies of scale for Norwegian, Danish and Finnish hospitals, but surprisingly, economies of scale for Swedish hospitals, which were actually aggregated to the county level.

8.1.2 Quality

Earlier studies show mixed results relating quality to productivity. Paper III adds to this with evidence of a slight trade-off between productivity and inpatient readmission, but no evidence of a trade-off between productivity and mortality. The latter result implies that it is possible to increase productivity while reducing mortality.

However, this is by no means a definitive conclusion on the relationship between quality and productivity, as quality of medical care has many dimensions (Donabedian 1966). Quality and health outcomes are intrinsically difficult to measure. What we have measured are merely *indicators* of quality, which do not necessarily present the optimal measure of quality. Our choice of quality indicators was guided not only by theory alone, but also by data availability and comparability. The indicators (readmission and out-of-hospital mortality) were chosen because we could create them based on registry data, and because we could apply the same indicator definitions in all Nordic countries. We acknowledge that these quality indicators do not exhaust the concept of quality, while at the same time, they do relate to quality. If two identical patients were to be treated at two different hospitals, we assume that higher quality treatment would be followed by a lower probability of readmission and death compared with lower quality treatment.

8.1.3 Activity-based funding (ABF) and hospital responses

ABF in Norway has been one of the greatest changes in the health care sector in the past few decades. However, there has not been much variation in issues such as the share of ABF. With small variability in one of the key instruments, it comes as no surprise that only small effects of ABF have been found (Kjerstad 2003; Yin et al. 2013). Overall, the health care system in Norway is egalitarian, and proper medical execution is probably more important to physicians than gaming the system. Biørn et al. (2010) found the response to the implementation of ABF to be heterogeneous amongst hospitals, as the best hospitals did not necessarily improve the most. This is also shown in Paper IV, where we see that while there was a price response, it was very low.

There is not enough variation within or between countries to analyse the effect of ABF share, especially within the short time frame in Papers II and III. An earlier study on the Nordic countries tested ABF as a second-stage explanatory variable, but found no effect (Kittelsen et

al. 2008). Paper I employ a longer time span, but if the share of ABF would have been used in Paper I, it would almost act as a dummy for some of the (early) years, as there would be no variation within a year and almost no variation between years. As for the comparative studies, ABF share would correlate perfectly with the country dummies, as there is no variation within the other countries or across time. Paper IV uses a novel approach by calculating ABF share as part of the incentive. We indirectly test the ABF share by testing how changes in the incentive affect coding. The argument is that both ABF share and price changes would have the same effect on hospitals. The analysis in Paper IV found no association on average between share of complicated patients and change in prices. There was, however, a positive association for medical DRGs and negative association for surgical DRGs, and there was a larger effect between the DRG pairs than within them.

Compared with U.S. studies (L. S. Dafny 2005; Silverman and Skinner 2004), recent European studies (Barros and Braun 2016; Januleviciute et al. 2016; Melberg et al. 2016) find similar price effects, but only of a small magnitude. Our study is in line with these findings. However, the operationalization of price incentive is different in Paper IV than in the other studies. Most studies looking at DRG pairs simply define the incentive as the spread in price between groups (L. S. Dafny 2005; Barros and Braun 2016; Januleviciute et al. 2016), but we have instead separated the effects of: 1) the level/size of the incentive, and 2) changes in the incentive. Januleviciute et al. (2016) undertake a similar decomposition when estimating overall price effects, but this is unrelated to DRG pair upcoding.

Pongpirul and Robinson (2013) argue that DRG creep and the different coding terms are often mixed. They categorize upcoding in three groups: 1) "a hospital coder may try to 'challenge code' by exploring the discharge summary to come up with the best possible codes" (Pongpirul and Robinson 2013), 2) "a hospital coder may go beyond the discharge summary and look for reimbursable conditions in the discharge summary and look for reimbursable conditions in the medical records. This is called DRG bracket creep, which is considered as a 'benign' form of upcoding" (Pongpirul and Robinson 2013), and finally, 3) add codes that the patient does not already have to increase income; this is by adding secondary diagnoses. Through retrospective statistical analyses on registry data, we cannot categorize the Norwegian changes in coding behaviour into these three categories; we can only speculate that it is likely a combination of all three. While in Paper IV, we saw that there was a large shift towards complicated cases, this

was not as linked to specific prices as feared. As discussed, the overall impact of ABF in Norway is large, but the measurable impact of its variance on secondary measures is low.

Using the same grouper for 16 years rests on the assumption that the use of diagnoses and procedures do not change in the period. Paper I assume this, and Paper IV tests this assumption. The results show that the use of secondary diagnoses, especially in paired DRGs, increase. Paper II attempted to minimize this specifically by averaging out the paired effect. This resulted in reducing the mean productivity level of Norwegian hospitals by two percentage points. Given this and the increase over time for secondary diagnoses, we must conclude that only a small part of the productivity growth in the period resulted from the increased use of secondary diagnoses. Similar results were also found by Biørn et al. (2003), where DRG creep (Simborg 1981) is briefly mentioned as they test efficiency by holding the case-mix index constant, which reduces the effect, indicating that a part of the productivity growth has been because of better coding. The size of this effect will probably depend on which years are included and at what level of analysis is undertaken. Petersen and Anthun (2008) found no hospital level effect on the use of secondary diagnoses on the case-mix index in the period 2002–2005.

Unsurprisingly, incentives work. A recent example from England demonstrates that a reimbursement increase for same-day surgery was followed by increase in volume, but no change in readmission or death rates (Allen et al. 2016). However, this can be interpreted as a price change that sped up a transition that would take place eventually. Payment by results may have consequences and has been criticized (Woolhandler et al. 2012). Norway uses more ABF than the other Nordic countries, has more secondary diagnoses and more readmissions. However, mortality is also lower, as shown in Paper III. Since DRGs incentivise a reduction in cost per stay, a fix may be to link payment to quality (Or and Häkkinen 2011). Given only part of the quality can be quantifiable, linking only parts of the output to payment may be a dangerous path (Woolhandler et al. 2012). For instance, some of the more easily quantifiable data are in-hospital deaths (as, for instance, used by Du et al. (2014); McKay and Deily (2008), and rewarding a low number of in-hospital deaths could be an incentive for the early discharge of dying patients. Out-of-hospital mortality is a thus a better quality indicator, but data on this might be more difficult to obtain. Early results from the English “Payment by Results” suggested that there was an increase in acute hospital activity, but not a decline in the quality of care, and no sign of increased out-of-hospital mortality replacing in-hospital mortality (Farrar et al. 2009). Another study found no long-term effect from pay for performance incentives on

30-day mortality (S. R. Kristensen et al. 2014). Many studies have also relied on complications as a quality measure (Hvenegaard et al. 2011; Kruse and Christensen 2013; McKay and Deily 2008). However, this might depend upon the honest registration of adverse events (such as hospital-acquired complications), which might not be incentivised.

8.2 Methodology

As we measure, errors may occur. The reliability (precision) of a measurement is how much random error is present. If much random error is present, precision is lowered; however, random error will not bias the average measurement we undertake. By increasing the number of observations, we can reduce the problem of random error; however, the measurement error and misclassification does not depend on sample size. High reliability is often considered a quality of good science. If measures are stable and consistent, we refer to it as high test-retest reliability.

However, validity (accuracy) relates not to random, but rather to systemic errors. Systematic errors will offset and bias results, and high validity will then imply that we actually measure what we set out to measure. Much of this thesis is based primarily on one data source (patient administrative data collected by the Norwegian Patient Registry), and we assume this source to be a valid measure of hospital activity.

Studies may have a selection bias if they are skewed in the data collection. The studies presented in this thesis are for the most part, based on complete datasets, which arguably do not have any selection bias. However, all the analysis is done on a population of hospital patients and hospital activity, so the knowledge gained from these studies cannot be applied to other populations. Paper IV use only a sample of the complete dataset. This could potentially be a biased sample, especially since this sample is a selection where we anticipate upcoding to take place. However, the purpose of the paper was to look at exactly how the incentive caused changes.

Upcoding is an elusive phenomenon and difficult to measure. If large-scale upcoding takes place, the activity data will seemingly show more severe patients or resource intensive activity than what is real. If so, the cost weights assigned to each episode will not correctly reflect the resources used by the hospitals for those episodes. In Paper IV, which examined upcoding, we did not find a very large shift, at least not large enough to describe the activities and cost weights

as being discordant. In Paper II, we also tested the impact of averaging the effects of DRG pairs. The results of this sensitivity analysis were that Norway's efficiency was slightly reduced, while the other Nordic countries remained unchanged. This indicates that first, upcoding is present, but more so in Norway, and second, that there is a small bias; however, the ranking of observations did not change much.

What constitutes a hospital may be increasingly hard to define, as there are a growing number of emergency care beds in municipal services, private hospitals and "district medical centres" with shared personnel resources. The problem is further accentuated in a comparative perspective. Papers II and III also utilize data from Sweden, Finland and Denmark, where definitional variances may cause systematic differences and lowered validity. The delineation between parts of the health system may be vastly different across countries. While there still may be definitional differences that we have not accounted for, we assume that also these data have negligible levels of measurement error. Nevertheless, the very act of international comparison opens some uncertainty, which we hope to resolve through thoroughness.

While most of the data were collected from the Norwegian Patient Registry and Statistics Norway, other sources have also supplied data. For Paper III municipal level data were collected from Statistics Norway, and quality data were either constructed from the activity data or collected from the *Folkeregisteret [National Register]*. These are registries with low probabilities of measurement error. Therefore, the issue may not be the actual data, but rather how they are used. In Paper IV, upcoding is a constructed operationalization because we cannot actually measure upcoding retrospectively. This opens the possibility for both random and systematic errors. We again rely on the source of data to be at least free of measurement error, but the constructed measurements may still include errors. However, our measures of upcoding are based on earlier studies (Barros and Braun 2016; L. S. Dafny 2005), but improved to capture better both the within- and between-effects of upcoding.

To conclude, I would argue that reliability is very high in these studies. The validity depends on the use of data, and for comparability, we depend on good definitions of the data and operationalization of the variables. All papers in this thesis are the result of laborious efforts to ensure comparability and reduce the level and effect of systematic errors.

8.3 Limitations

It has been recommended that studies separate their analysis between health policy actors on the one hand, and health care systems on the other (Marmor and Wendt 2012); mixing these could be problematic. We acknowledge that there are relevant and interesting perspectives and insights to be gained by studying actors and institutions. In addition, most of the data are at the micro level. However, in this thesis, the focus is deliberately on the health care system, not the actors. Our perspective is therefore not what individual actors do, but rather how regulations and health policies affect the general system. Jacobs et al. (2006) observe that the unit of analysis should capture the entire production process, should have discretion about the technological process of converting inputs to outputs and should be comparable. In all papers in this thesis, the hospitals are the level of study, but the perspective is on how events at that level sum up the health system as a whole.

This thesis is based upon a quantitative approach and utilizes registry data. Productivity is a phenomenon that happens at the hospital and in the wards; nevertheless a qualitative approach is not the obvious choice for methodology. A qualitative approach will ask different questions and consequently arrive at different answers than a quantitative approach. There is, however, one exception, as Paper IV examines upcoding, which one could argue is impossible to research quantitatively. As we only have data from one source (patient administrative systems in hospitals), we cannot detect true upcoding. Journal revision may be the only possible method that can retrospectively say something about *real* as opposed to *upcoded* diagnoses (Jørgenvåg and Hope 2005). Journal revision is, however, not possible because of our lack of medical training, and, more importantly, the method does not scale well compared with the 3.18 million case dataset used in Paper IV. However, even journal revision is far detached from the actual process of diagnostic coding. Upcoding can be considered a qualitative phenomenon that takes place in a meeting between a patient and a physician in the presence of norms, rules, regulations and incentives. To sort out each incidence of upcoding requires participation in the actual coding processes. The aim here has instead been to analyse the systemic relationship between upcoding and prices. Being trained quantitatively and skilled in the use of registry data, I have chosen to resort to the only tools I can use: statistics for large datasets.

For the productivity analyses in Papers I, II and III, we rely on the well-used estimating technique of DEA. O'Neill et al. (2008) point out that in the economics literature, SFA is perceived to be a better method than DEA. However, the ease of decomposing inefficiencies is important in hospital efficiency studies, where SFA has some drawbacks, according to O'Neill et al. (2008). There have been some comparisons of SFA and DEA, and although not identical, the results seem to correspond. In Paper II, we found that the results were robust to the choice of either SFA or bootstrapped DEA. However, this does not mean that it is not important which methods are used.

Another limitation of these studies is related to hospital inputs. Capital costs are an important part of hospital costs not accounted for by this study. Estimates of hospital productivity, including capital costs, show that including capital costs as a separate input dimension increases productivity for the years where information on capital is available (Anthun et al. 2016). However, other inputs might also be relevant. Recently a study of physician productivity showed a productivity decline over time (Johannessen et al. 2017). While the output measures in our studies have been thoroughly case-mix adjusted both within and across years, we rely on the deflator (in Table 2 above on page 53) to be accurate when adjusting input. The cumulative effects of price changes over 16 years are significant and may have a substantial impact on the slope and level of the productivity development.

Most journals have strict limitations on the size of published articles, and all papers have many more related analyses that could have been performed. Paper I would benefit greatly from a second-stage analysis that could explain the level and development over such a long period. Similarly, Paper IV would benefit from exploiting the individual-level data better.

In medicine and other branches of science, publishing biases may have a large impact, and it is important to counteract these effects. In these studies, it is probably of less importance, and while there may be some ideas and analyses that remain unpublished (Paper IV undertook a selection of the years to be included due to lack of comparability across years, and similarly, capital costs were not included in Paper I), all four papers paint a picture of mixed results and small effects. In a worst-case scenario, we can expect very publishing bias-prone studies to have clearer and stronger results. However, mixed results and small effects alone will not exonerate the researcher. We declare that none of the research in this study is wilfully unpublished. The funding of this study was provided by a neutral research body that is not hoping for a result in any specific direction.

8.4 Further research

Paper II, which is a follow-up of Kittelsen et al. (2008) and Kittelsen et al. (2009) and a precursor to Paper III, established a country-specific frontier. These represent institutional settings and national policies that are out of the control of hospitals and specific to each country. Papers II and III have tested some potential explanatory factors regarding what may cause these differences in productivity, and more research should attempt to uncover the substantial constraints that these impose on hospitals. Are they policies that can change, related to delineation of services and institutions, or is the answer possible to find by testing other hospital inputs such as personnel, beds, salary overheads, research and teaching costs?

Quality of care and health outcomes have long been important issues in studies of reforms, financing and productivity. However, there has been many mixed results, indicating that quality and health outcomes are elusive and difficult to measure. The quality link in productivity analyses has been shown to be important, but the trade-off has not been as large as expected. Quality has also been included as a part of the funding scheme in some countries, for instance, payment by results. Quality is also related to upcoding. Many of the studies on health outcomes have been cross-sectional studies at the hospital level, and we propose to examine this further by either exploiting individual-level data more or using a longitudinal perspective. Paper I established both the usefulness and plausibility of such a perspective, and Paper III demonstrated how patient register data could provide valuable insights into quality.

Most studies that look at ABF or upcoding, at least in a Norwegian context, find only small effects or present regressions with low explanatory power. More research should go into including substantial explanations into regressions. This could make better use of individual-level data, or by collecting more data at the hospital level, explain hospital differences such as productivity or local use of ABF. In this thesis, very large datasets were used to respond to these questions. Using different methods would yield different answers to different questions. As upcoding is a qualitative phenomenon, it may also be appropriate to study it using other methods.

8.5 Concluding remarks

What is productivity and productivity growth? Is it a question of the personnel working faster and doing more? Donaldson and Magnussen (1992) claim that inefficiency is the choice of hospitals, i.e., they have discretion on how much slack there should be at any time since "…hospital inefficiencies arise because hospitals want them to". In the real world, the situation is arguably more complex. Should hospitals be expected to be productive in an economic sense? Should hospitals be input minimizing? Tjerbo and Hagen show that additional funding and deficit bailouts were provided by the Norwegian parliamentary politicians with the hope that this would provide good results in elections (Tjerbo and Hagen 2009). Against this and the rising wealth of Norway, it has not been easy to argue politically against growth in health expenditure. However, the situation with deficits ended after hospitals faced increasingly hard budget constraints. In later years, hospitals have operated at balance or with a surplus. Hospitals in Norway do not create a profit for their owner, but to increase investment budgets, they must operate with a surplus, and the productivity growth after the reform contributed to this improvement.

The foundation for the Norwegian hospital policy is the National Health and Hospital plan, which assumes an annual one per cent input saving (less personnel intensive) productivity growth for future hospital services (Stortingsforhandling 2015). In Paper I, we estimated an average productivity growth of 1.5 per cent, however in the years after 2002 it has been below 0.5 per cent. As researchers asking questions about the productivity of others, there may well be ethical considerations. Employees working at hospitals are already under continuous pressure to increase productivity, and research on productivity will indirectly label a lack of efficiency as inefficiency and slack. However, throughout this thesis, we have adopted a health system perspective, even when using individual- or hospital-level data, the goal being to say something about how the governance and system rules affect the overall system. By studying productivity, the purpose is not for increased productivity per se to become a goal of a health system. Hospitals should improve productivity if the slack is not beneficial to patients or personnel, i.e., if productivity can be increased without a detrimental effect on patients and personnel, such that productivity is a valuable improvement in a health system perspective. This is especially true if we remember that the overall goal is not to perform as many operations and treatments as possible, but to improve health.

The central theme of this thesis is ABF and productivity growth. We have studied how productivity developed in Norway from 1999 to 2014. In the four papers, we have decomposed and explained Norwegian developments and compared Norway to other Nordic countries. Finally, we have shown how a small part of this productivity growth owes to upcoding. This thesis has shown the importance of a longitudinal perspective (Papers I and IV) and a comparative approach (Papers II and III). The major strength of this thesis is a methodological point common to all four papers. We have shown the importance of sorting out differences between cross-sectional differences and differences over time. In Paper I, we saw both the overall development and the decomposition of causes for year-to-year changes. Similarly, in Papers II and III, instead of being limited to a cross-section of hospitals within a country, we compare and decompose trends over time and between countries. Paper IV also shows development in the long run and separates the incentive effect into *between* and *within* effects, thus studying both the level of and changes in the incentives.

As this is a quantitative thesis, it shows only the association, and therefore cannot discuss issues such as meaning and intent, which are important to suggest good and efficient health policies. By improving the methods for hospital comparisons, we have shown the importance of a longitudinal perspective (Papers I and IV) and a comparative approach (Papers II and III), and through methodical thoroughness, this thesis presents four papers that add valuable insights to the literature.

9 Literature

Ahgren, B. (2008). Is it better to be big?: The reconfiguration of 21st century hospitals: Responses to a hospital merger in Sweden. *Health Policy, 87*(1), 92-99.

Aiken, L. H., Sloane, D. M., Bruyneel, L., Van den Heede, K., Griffiths, P., Busse, R., et al. (2014). Nurse staffing and education and hospital mortality in nine European countries: a retrospective observational study. *The Lancet, 383*(9931), 1824-1830.

Aletras, V. H. (1999). A comparison of hospital scale effects in short-run and long-run cost functions. *Health Economics, 8*(6), 521-530.

Allen, T., Fichera, E., & Sutton, M. (2016). Can payers use prices to improve quality? Evidence from English hospitals. *Health Economics, 25*(1), 56-70.

Anthun, K. S., Goude, F., Häkkinen, U., Kittelsen, S., Kruse, M., Medin, E., et al. (2013). Eurohope hospital level analysis: material, methods and indicators. *Eurohope discussion papers No 10*. Helsinki: THL.

Anthun, K. S., Kittelsen, S. A., & Magnussen, J. (2016). Produktivitet i spesialisthelsetjenesten. University of Oslo, Health Economics Research Programme Working paper 2016: 7.

Asmild, M., Hollingsworth, B., & Birch, S. (2013). The scale of hospital production in different settings: One size does not fit all. *Journal of Productivity Analysis, 40*(2), 197-206.

Averill, R. (1991). Development. In R. Fetter, D. Brand, & D. Gamanch (Eds.), *DRGs Their Design and Development*: Health Administration Press, Ann Arbor

Banker, R. D. (1984). Estimating most productive scale size using data envelopment analysis. *European Journal of Operational Research, 17*(1), 35-44.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science, 30*(9), 1078-1092.

Barer, M. L. (1982). Case mix adjustment in hospital cost analysis: Information theory revisited. *Journal of Health Economics, 1*(1), 53-80, doi:10.1016/0167-6296(82)90021-2.

Barros, P., & Braun, G. (2016). Upcoding in a National Health Service: the evidence from Portugal. [Research Article]. *Health Economics, 26*(5), 600-618, doi:10.1002/hec.3335.

Battese, G. E., & Coelli, T. J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical economics, 20*(2), 325-332.

Berlowitz, D. R., Ash, A. S., Brandeis, G. H., & et al. (1996). Rating long-term care facilities on pressure ulcer development: Importance of case-mix adjustment. *Annals of Internal Medicine, 124*(6), 557-563, doi:10.7326/0003-4819-124-6-199603150-00003.

Berta, P., Callea, G., Martini, G., & Vittadini, G. (2010). The effects of upcoding, cream skimming and readmissions on the Italian hospitals efficiency: a population-based investigation. *Economic Modelling, 27*(4), 812-821.

Biørn, E., Hagen, T. P., Iversen, T., & 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*(4), 271-283.

Biørn, E., Hagen, T. P., Iversen, T., & Magnussen, J. (2010). How different are hospitals' responses to a financial reform? The impact on efficiency of activity-based financing. *Health care management science, 13*(1), 1-16.

Bowblis, J. R., & Brunt, C. S. (2014). Medicare skilled nursing facility reimbursement and upcoding. *Health Economics, 23*(7), 821-840.

Burgess Jr, J. F., & Wilson, P. W. (1995). Decomposing hospital productivity changes, 1985–1988: a nonparametric Malmquist approach. *Journal of Productivity Analysis, 6*(4), 343-363.

Burns, E. M., Rigby, E., Mamidanna, R., Bottle, A., Aylin, P., Ziprin, P., et al. (2012). Systematic review of discharge coding accuracy. *Journal of Public Health, 34*(1), 138-148.

Busse, R. (2012). Do Diagnosis-Related Groups Explain Variations in Hospital Costs and Length of Stay? - Analyses from the EuroDRG Project for 10 Episodes of Care across 10 European Countries. [Editorial]. *Health Economics, 21*, 1-5, doi:10.1002/hec.2861.

Campbell, S. E., Campbell, M. K., Grimshaw, J. M., & Walker, A. E. (2001). A systematic review of discharge coding accuracy. *Journal of Public Health, 23*(3), 205-211.

Carter, G. M., & Ginsburg, P. B. (1985). The medicare case mix index increase: Medical Practice Changes, Aging and DRG Creep. *Rand Publication Series*. Santa Monica: Rand Corporation Report R-3292-HCFA.

Carter, G. M., Newhouse, J. P., & Relles, D. A. (1990). How much change in the case mix index is DRG creep? *Journal of Health Economics, 9*(4), 411-428.

Castelli, A., Street, A., Verzulli, R., & Ward, P. (2015). Examining variations in hospital productivity in the English NHS *The European journal of health economics, 16*(3), 243-254.

Caves, D. W., Christensen, L. R., & Diewert, W. E. (1982). The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica: Journal of the Econometric Society*, 1393-1414.

Charlson, M. E., Pompei, P., Ales, K. L., & MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases, 40*(5), 373-383.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6), 429-444.

Chilingerian, J. A. (1994). Exploring why some physicians' hospital practices are more efficient: taking DEA inside the hospital. In A. Charnes, W. W. Cooper, A. Y. Lewin, & L. M. Seiford (Eds.), *Data Envelopment Analysis: Theory, Methodology, and Applications* (pp. 167-193): Springer.

Chowdhury, H., Wodchis, W., & Laporte, A. (2011). Efficiency and technological change in health care services in Ontario: An application of Malmquist Productivity Index with bootstrapping. *International Journal of Productivity and Performance Management, 60*(7), 721-745.

Chowdhury, H., Zelenyuk, V., Laporte, A., & Wodchis, W. P. (2014). Analysis of productivity, efficiency and technological changes in hospital services in Ontario: How does case-mix matter? *International Journal of Production Economics, 150*, 74-82.

Clement, J. P., Valdmanis, V. G., Bazzoli, G. J., Zhao, M., & Chukmaitov, A. (2008). Is more better? An analysis of hospital outcomes and efficiency with a DEA model of output congestion. *Health care management science, 11*(1), 67-77.

Cots, F., Chiarello, P., Salvador, X., Castells, X., & Quentin, W. (2011). DRG-based hospital payment: Intended and unintended consequences. In *Diagnosis-Related Groups in Europe: Moving towards transparency, efficiency and quality in hospitals* (pp. 75-92).

Dafny, L., & Dranove, D. (2009). Regulatory exploitation and management changes: upcoding in the hospital industry. *Journal of Law and Economics, 52*(2), 223-250.

Dafny, L. S. (2005). How Do Hospitals Respond to Price Changes? *The American Economic Review, 95*(5), 1525-1547.

Delhey, J., & Newton, K. (2005). Predicting cross-national levels of social trust: global pattern or Nordic exceptionalism? *European Sociological Review, 21*(4), 311-327.

Dervaux, B., Ferrier, G. D., Leleu, H., & Valdmanis, V. (2004). Comparing French and US hospital technologies: a directional input distance function approach. *Applied Economics, 36*(10), 1065-1081.

Donabedian, A. (1966). Evaluating the quality of medical care. *The Milbank memorial fund quarterly, 44*(3), 166-206.

Donaldson, C., & Magnussen, J. (1992). DRGs: the road to hospital efficiency. *Health Policy, 21*(1), 47-64.

Dranove, D. (1998). Economies of scale in non-revenue producing cost centers: implications for hospital mergers. *Journal of Health Economics, 17*(1), 69-83.

Drösler, S. (2008). Facilitating cross national comparisons of indicators for patient safety at the health system level in the OECD countries. *OECD Health Tehcnical Papers: 19*. Paris: OECD.

Du, J., Wang, J., Chen, Y., Chou, S.-Y., & Zhu, J. (2014). Incorporating health outcomes in Pennsylvania hospital efficiency: an additive super-efficiency DEA approach. *Annals of Operations Research, 221*(1), 161-172.

Eijkenaar, F., Emmert, M., Scheppach, M., & Schöffski, O. (2013). Effects of pay for performance in health care: a systematic review of systematic reviews. *Health Policy, 110*(2), 115-130.

Ellis, R. P., & McGuire, T. G. (1986). Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of Health Economics, 5*(2), 129-151.

Farrar, S., Yi, D., Sutton, M., Chalkley, M., Sussex, J., & Scott, A. (2009). Has payment by results affected the way that English hospitals provide care? Difference-in-differences analysis. *Bmj, 339*, b3047, doi:10.1136/bmj.b3047.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General), 120*(3), 253-290.

Ferrier, G. D., & Trivitt, J. S. (2013). Incorporating quality into the measurement of hospital efficiency: a double DEA approach. *Journal of Productivity Analysis, 40*(3), 337-355.

Fisher, E. S., Whaley, F. S., Krushat, W. M., Malenka, D. J., Fleming, C., Baron, J. A., et al. (1992). The accuracy of Medicare's hospital claims data: progress has been made, but problems remain. *American Journal of Public Health, 82*(2), 243-248.

Flodgren, G., Eccles, M. P., Shepperd, S., Scott, A., Parmelli, E., & Beyer, F. R. (2011). An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes. *Cochrane Database Syst Rev, 7*(7).

Färe, R., Grosskopf, S., Lindgren, B., & Poullier, J.-P. (1997). Productivity growth in health-care delivery. *Medical Care, 35*(4), 354-366.

Färe, R., Grosskopf, S., Lindgren, B., & Roos, P. (1992). Productivity changes in Swedish pharamacies 1980–1989: A non-parametric Malmquist approach. In T. R. J. Gulledge, & C. A. Know Lovell (Eds.), *International Applications of Productivity and Efficiency Analysis* (pp. 81-97): Springer.

Färe, R., Grosskopf, S., Lindgren, B., & Roos, P. (1994a). Productivity developments in Swedish hospitals: a Malmquist output index approach. In A. Charnes, W. W. Cooper, A. Y. Lewin, & L. M. Seiford (Eds.), *Data envelopment analysis: theory, methodology, and applications* (pp. 253-272): Springer.

Färe, R., Grosskopf, S., Norris, M., & Zhang, Z. (1994b). Productivity growth, technical progress, and efficiency change in industrialized countries. *The American Economic Review*, 66-83.

Førsund, F. R., & Hjalmarsson, L. (2004). Are all scales optimal in DEA? Theory and empirical evidence. *Journal of Productivity Analysis, 21*(1), 25-48.

Geissler, A., Quentin, W., Scheller-Kreinsen, D., & Busse, R. (2011). Introduction to DRGs in Europe: Common objectives across different hospital systems. In R. Busse, A. Geissler, W. Quentin, & M. Wiley (Eds.), *Diagnosis Related Groups in Europe: Moving Towards Transparency, Efficiency and Quality in Hospitals* (pp. 9-21).

Gerdtham, U.-G., Rehnberg, C., & Tambour, M. (1999). The impact of internal markets on health care efficiency: evidence from health care reforms in Sweden. *Applied Economics, 31*(8), 935-945.

Given, R. S. (1996). Economies of scale and scope as an explanation of merger and output diversification activities in the health maintenance organization industry. *Journal of Health Economics, 15*(6), 685-713.

Grosskopf, S., & Valdmanis, V. (1993). Evaluating hospital performance with case-mix-adjusted outputs. *Medical Care, 31*(6), 525-532.

Gutacker, N., Bojke, C., Daidone, S., Devlin, N. J., Parkin, D., & Street, A. (2013). Truly inefficient or providing better quality of care? Analysing the relationship between risk-adjusted hospital costs and patients' health outcomes. *Health Economics, 22*(8), 931-947.

Hagen, T. P., & Kaarbøe, O. M. (2006). The Norwegian hospital reform of 2002: central government takes over ownership of public hospitals. *Health Policy, 76*(3), 320-333.

Hagen, T. P., Veenstra, M., & Stavem, K. (2006). Efficiency and patient satisfaction in Norwegian hospitals. *Health Organization Research Norway Working Paper 2006:1*: University of Oslo / SINTEF.

Halkos, G. E., & Tzeremes, N. G. (2011). A conditional nonparametric analysis for measuring the efficiency of regional public healthcare delivery: An application to Greek prefectures. *Health Policy, 103*(1), 73-82.

Halsteinli, V., Kittelsen, S. A., & Magnussen, J. (2010). Productivity growth in outpatient child and adolescent mental health services: the impact of case-mix adjustment. *Social science & medicine, 70*(3), 439-446.

Hollingsworth, B. (2003). Non-parametric and parametric applications measuring efficiency in health care. *Health care management science, 6*(4), 203-218.

Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics, 17*(10), 1107-1128.

Hollingsworth, B., & Parkin, D. (2001). The efficiency of the delivery of neonatal care in the UK. *Journal of Public Health, 23*(1), 47-50.

Hollingsworth, B., & Wildman, J. (2003). The efficiency of health production: re-estimating the WHO panel data using parametric and non-parametric approaches to provide additional information. *Health Economics, 12*(6), 493-504.

Hsia, D. C., Krushat, W. M., Fagan, A. B., Tebbutt, J. A., & Kusserow, R. P. (1988). Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. *The New England Journal of Medicine, 318*(6), 352-355.

Hsing, Y., & Bond, E. O. (1995). In search of optimal productivity and hospital size: A case study. *The Health Care Manager, 14*(2), 50-55.

Hussey, P. S., Wertheimer, S., & Mehrotra, A. (2013). The association between health care quality and cost: a systematic review. *Annals of Internal Medicine, 158*(1), 27-34.

Hvenegaard, A., Arendt, J. N., Street, A., & Gyrd-Hansen, D. (2011). Exploring the relationship between costs and quality: Does the joint evaluation of costs and quality alter the ranking of Danish hospital departments? *The European journal of health economics, 12*(6), 541-551.

Hvenegaard, A., Street, A., Sørensen, T. H., & Gyrd-Hansen, D. (2009). Comparing hospital costs: What is gained by accounting for more than a case-mix index? *Social science & medicine, 69*(4), 640-647, doi:10.1016/j.socscimed.2009.05.047.

Häkkinen, U., Iversen, T., Peltola, M., Rehnberg, C., & Seppälä, T. T. (2015). Towards Explaining International Differences in Health Care Performance: Results of the EuroHOPE Project. *Health Economics, 24*, 1-4, doi:10.1002/hec.3282.

Jacobs, R. (2001). Alternative Methods to Examine Hospital Efficiency: Data Envelopment Analysis and Stochastic Frontier Analysis. *Health care management science, 4*(2), 103-115, doi:10.1023/a:1011453526849.

Jacobs, R., Smith, P. C., & Street, A. (2006). *Measuring efficiency in health care: analytic techniques and health policy*: Cambridge University Press.

Jakobsen, M. L. F. (2010). The Effects of New Public Management: Activity-based Reimbursement and Efficiency in the Scandinavian Hospital Sectors. *Scandinavian Political Studies, 33*(2), 113-134.

Januleviciute, J., Askildsen, J. E., Kaarboe, O., Siciliani, L., & Sutton, M. (2016). How do Hospitals Respond to Price Changes? Evidence from Norway. *Health Economics, 25*, 620-636, doi:10.1002/hec.3179.

Jegers, M., Kesteloot, K., De Graeve, D., & Gilles, W. (2002). A typology for provider payment systems in health care. *Health Policy, 60*(3), 255-273.

Johannessen, K. A., Kittelsen, S. A. C., & Hagen, T. P. (2017). Assessing physician productivity following Norwegian hospital reform: A panel and data envelopment analysis. *Social science & medicine*, doi:10.1016/j.socscimed.2017.01.008.

Jørgenvåg, R., & Hope, Ø. B. (2005). Kvalitet på medisinsk koding og ISF-refusjoner. I hvilken grad er journalgjennomgang et nyttig verktøy [Quality of diagnostic coding and activity based financing. To what extent is journal revision a useful tool?]. SINTEF Report STF78 A055501, Trondheim Norway.

Kalseth, B., Anthun, K. S., Hope, Ø., Kittelsen, S. A., & Persson, B. A. (2011). Spesialisthelsetjenesten i Norden. Sykehusstruktur, styringsstruktur og lokal arbeidsorganisering som mulig forklaring på kostnadsforskjeller mellom landene. Trondheim: SINTEF A19615.

Kittelsen, S. A., Anthun, K. S., Kalseth, B., Kalseth, J., Halsteinli, V., & Magnussen, J. (2009). En komparativ analyse av spesialisthelsetjenesten i Finland, Sverige, Danmark og Norge: Aktivitet, ressursbruk og produktivitet 2005-2007. Trondheim: SINTEF A12200.

Kittelsen, S. A., Magnussen, J., & Anthun, K. S. (2007). Sykehusproduktivitet etter statlig overtakelse: En nordisk komparativ analyse. *Working paper 2007:1*. Oslo: University of Oslo, Health Economics Research Programme.

Kittelsen, S. A., Magnussen, J., Anthun, K. S., Häkkinen, U., Linna, M., Medin, E., et al. (2008). *Hospital productivity and the Norwegian ownership reform: A Nordic comparative study* (Stakes Discussion Papers 3). Helsinki: Stakes.

Kjekshus, L., & Hagen, T. (2007). Do hospital mergers increase hospital efficiency? Evidence from a National Health Service country. *Journal of health services research & policy, 12*(4), 230-235.

Kjerstad, E. (2003). Prospective Funding of General Hospitals in Norway—Incentives for Higher Production? *International Journal of Health Care Finance and Economics, 3*(4), 231-251, doi:10.1023/a:1026084304382.

Kristensen, S. R. (2016). Financial Penalties for Performance in Health Care. [Editorial]. *Health Economics, 26*, 143-148, doi:10.1002/hec.3463.

Kristensen, S. R., Meacock, R., Turner, A. J., Boaden, R., McDonald, R., Roland, M., et al. (2014). Long-term effect of hospital pay for performance on mortality in England. *New England Journal of Medicine, 371*(6), 540-548.

Kristensen, T., Bogetoft, P., & Pedersen, K. M. (2010). Potential gains from hospital mergers in Denmark. *Health care management science, 13*(4), 334-345.

Kristensen, T., Olsen, K. R., Kilsmark, J., Lauridsen, J. T., & Pedersen, K. M. (2012). Economies of scale and scope in the Danish hospital sector prior to radical restructuring plans. *Health Policy, 106*(2), 120-126.

Kristensen, T., Olsen, K. R., Kilsmark, J., & Pedersen, K. M. (2008). *Economies of scale and optimal size of hospitals: Empirical results for Danish public hospitals*: Syddansk Universitet.

Kruse, M., & Christensen, J. (2013). Is quality costly? Patient and hospital cost drivers in vascular surgery. *Health economics review, 3*(1), 1.

Kuhn, M., & Siciliani, L. (2008). Upcoding and optimal auditing in health care (or the economics of DRG creep). *CEPR Discussion Paper No. DP6689*.

Liang, L. L. (2015). Do Diagnosis-Related Group-Based Payments Incentivise Hospitals to Adjust Output Mix? *Health Economics, 24*(4), 454-469.

Linna, M., & Häkkinen, U. (1999). Determinants of Cost efficiency of Finnish Hospitals: A Comparison of DEA and SFA. *Helsinki University of Technology: Systems Analysis Laboratory Research Report #A78*.

Linna, M., Häkkinen, U., & Magnussen, J. (2006). Comparing hospital cost efficiency between Norway and Finland. *Health Policy, 77*(3), 268-278.

Linna, M., Häkkinen, U., Peltola, M., Magnussen, J., Anthun, K. S., Kittelsen, S., et al. (2010). Measuring cost efficiency in the Nordic Hospitals—a cross-sectional comparison of public hospitals in 2002. *Health care management science, 13*(4), 346-357.

Lovell, C. K. (2003). The decomposition of Malmquist productivity indexes. *Journal of Productivity Analysis, 20*(3), 437-458.

Lyttkens, C. H., Christiansen, T., Häkkinen, U., Kaarboe, O., Sutton, M., & Welander, A. (2016). The core of the Nordic health care system is not empty. *Nordic Journal of Health Economics, 4*(1), pp. 7-27.

Magnussen, J. (1996). Efficiency measurement and the operationalization of hospital production. *Health services research, 31*(1), 21.

Magnussen, J., Hagen, T. P., & Kaarboe, O. M. (2007). Centralized or decentralized? A case study of Norwegian hospital reform. *Social science & medicine, 64*(10), 2129-2137.

Magnussen, J., & Solstad, K. (1994). Case-based hospital financing: the case of Norway. *Health Policy, 28*(1), 23-36.

Maniadakis, N., & Thanassoulis, E. (2004). A cost Malmquist productivity index. *European Journal of Operational Research, 154*(2), 396-409.

Marini, G., & Miraldo, M. (2009). Economies of scale and scope in the English hospital sector. Imperial College London, Business School. Discussion paper 2009/05.

Marmor, T., & Wendt, C. (2012). Conceptual frameworks for comparing healthcare politics and policy. *Health Policy, 107*(1), 11-20.

Martinussen, P. E., & Hagen, T. P. (2009). Reimbursement systems, organisational forms and patient selection: evidence from day surgery in Norway. *Health Economics, Policy and Law, 4*(02), 139-158.

Martinussen, P. E., & Midttun, L. (2004). Day surgery and hospital efficiency: empirical analysis of Norwegian hospitals, 1999-2001. *Health Policy, 68*(2), 183-196, doi:10.1016/j.healthpol.2003.09.003.

McKay, N. L., & Deily, M. E. (2008). Cost inefficiency and hospital health outcomes. *Health Economics, 17*(7), 833-848.

Medin, E., Anthun, K. S., Häkkinen, U., Kittelsen, S. A., Linna, M., Magnussen, J., et al. (2011). Cost efficiency of university hospitals in the Nordic countries: a cross-country analysis. *The European journal of health economics, 12*(6), 509-519.

Medin, E., Häkkinen, U., Linna, M., Anthun, K. S., Kittelsen, S. A., Rehnberg, C., et al. (2013). International hospital productivity comparison: experiences from the Nordic countries. *Health Policy, 112*(1), 80-87.

Melberg, H. O., Beck Olsen, C., & Pedersen, K. (2016). Did hospitals respond to changes in weights of Diagnosis Related Groups in Norway between 2006 and 2013? *Health Policy, 120*(9), 992-1000, doi:10.1016/j.healthpol.2016.07.013.

Milstein, R., & Schreyoegg, J. (2016). Pay for performance in the inpatient sector: A review of 34 P4P programs in 14 OECD countries. *Health Policy, 120*(10), 1125-1140, doi:10.1016/j.healthpol.2016.08.009.

Mobley, L. R. I., & Magnussen, J. (1998). An international comparison of hospital efficiency: does institutional environment matter? *Applied Economics, 30*(8), 1089-1100.

Mortimer, D. (2002). Competing methods for efficiency measurement: a systematic review of direct DEA vs SFA/DFA comparisons. Centre for Health Program Evaluation, Monash University, Australia. Working Paper 136.

Nayar, P., & Ozcan, Y. A. (2008). Data envelopment analysis comparison of hospital efficiency and quality. *Journal of medical systems, 32*(3), 193-199.

Neby, S., Lægreid, P., Mattei, P., & Feiler, T. (2015). Bending the Rules to Play the Game: Accountability, DRG and Waiting List Scandals in Norway and Germany. *European Policy Analysis, 1*(1), 127-148.

O'Reilly, J., Busse, R., Häkkinen, U., Or, Z., Street, A., & Wiley, M. (2012). Paying for hospital care: the experience with implementing activity-based funding in five European countries. *Health Economics, Policy and Law, 7*(Special Issue 01), 73-101, doi:10.1017/S1744133111000314.

O'Neill, L., Rauner, M., Heidenberger, K., & Kraus, M. (2008). A cross-national comparison and taxonomy of DEA-based hospital efficiency studies. *Socio-Economic Planning Sciences, 42*(3), 158-189.

Ogundeji, Y. K., Bland, J. M., & Sheldon, T. A. (2016). The effectiveness of payment for performance in health care: A meta-analysis and exploration of variation in outcomes. *Health Policy, 120*(10), 1141-1150.

Or, Z., & Häkkinen, U. (2011). DRGs and quality: for better or worse. In R. Busse, A. Geissler, W. Quentin, & M. Wiley (Eds.), *Diagnosis-Related Groups in Europe: Moving towards transparency, efficiency and quality in hospitals* (pp. 115-129). Maidenhead: Open University Press.

Palmer, K. S., Agoritsas, T., Martin, D., Scott, T., Mulla, S. M., Miller, A. P., et al. (2014). Activity-Based Funding of Hospitals and Its Impact on Mortality, Readmission,

Discharge Destination, Severity of Illness, and Volume of Care: A Systematic Review and Meta-Analysis. *PLoS ONE, 9*(10), e109975, doi:10.1371/journal.pone.0109975.

Pedersen, M., Kalseth, B., Lilleeng, S. E., Mehus, K. H., Pedersen, P. B., & Sitter, M. (2016). Private aktører i spesialisthelsetjenesten. Omfang og utvikling 2010-2014. Oslo: Helsedirektoratet IS-2450.

Petersen, S. Ø., & Anthun, K. S. (2008). Endringer i DRG-indeks. Beskrivelse av metode og resultater 2002-2005 [Changes in the DRG-index. Description of methods and results 2002-2005]. Trondheim: SINTEF Helse Report A1369.

Pongpirul, K., & Robinson, C. (2013). Hospital manipulations in the DRG system: a systematic scoping review. *Asian Biomedicine, 7*, 301-310.

Posnett, J. (1999). Is bigger better? Concentration in the provision of secondary care. *British medical journal, 319*(7216), 1063.

Preyra, C. (2004). Coding response to a case-mix measurement system based on multiple diagnoses. *Health services research, 39*(4p1), 1027-1046.

Preyra, C., & Pink, G. (2006). Scale and scope efficiencies through hospital consolidations. *Journal of Health Economics, 25*(6), 1049-1068.

Puig-Junoy, J. (1998). Measuring health production performance in the OECD. *Applied Economics Letters, 5*(4), 255-259.

Retzlaff-Roberts, D., Chang, C. F., & Rubin, R. M. (2004). Technical efficiency in the use of health care resources: a comparison of OECD countries. *Health Policy, 69*(1), 55-72.

Rosenberg, M. A., & Browne, M. J. (2001). The Impact of the Inpatient Prospective Payment System and Diagnosis-Related Groups. *North American Actuarial Journal, 5*(4), 84-94.

Rutledge, R. W., Parsons, S., & Knaebel, R. (1995). Assessing hospital efficiency over time: an empirical application of data envelopment analysis. *Journal of Information Technology Management, 6*, 13-24.

Salem-Schatz, S., Moore, G., Rucker, M., & Pearson, S. D. (1994). The case for case-mix adjustment in practice profiling: When good apples look bad. *JAMA, 272*(11), 871-874, doi:10.1001/jama.1994.03520110051028.

Samdata (2015). SAMDATA Spesialisthelsetjenesten [Norwegian annual statistics on specialized hospital care]. Oslo: Helsedirektoratet IS-2348.

Sax, H., Pittet, D., & and the Swiss, N. N. (2002). Interhospital differences in nosocomial infection rates: Importance of case-mix adjustment. *Archives of Internal Medicine, 162*(21), 2437-2442, doi:10.1001/archinte.162.21.2437.

Serdén, L., Lindqvist, R., & Rosén, M. (2003). Have DRG-based prospective payment systems influenced the number of secondary diagnoses in health care administrative data? *Health Policy, 65*(2), 101-107.

Silverman, E., & Skinner, J. (2004). Medicare upcoding and hospital ownership. *Journal of Health Economics, 23*(2), 369-389.

Simar, L., & Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management science, 44*(1), 49-61.

Simar, L., & Wilson, P. W. (2000). Statistical inference in nonparametric frontier models: The state of the art. *Journal of Productivity Analysis, 13*(1), 49-78.

Simborg, D. W. (1981). DRG creep: a new hospital-acquired disease. *The New England Journal of Medicine, 304*(26), 1602.

Sommersguter-Reichmann, M. (2000). The impact of the Austrian hospital financing reform on hospital productivity: empirical evidence on efficiency and technology changes using a non-parametric input-based Malmquist approach. *Health care management science, 3*(4), 309-321.

Spinks, J., & Hollingsworth, B. (2009). Cross-country comparisons of technical efficiency of health production: a demonstration of pitfalls. *Applied Economics, 41*(4), 417-427.

Steinbusch, P. J., Oostenbrink, J. B., Zuurbier, J. J., & Schaepkens, F. J. (2007). The risk of upcoding in casemix systems: a comparative study. *Health Policy, 81*(2), 289-299.

Steinmann, L., Dittrich, G., Karmann, A., & Zweifel, P. (2004). Measuring and comparing the (in) efficiency of German and Swiss hospitals. *The European Journal of Health Economics, formerly: HEPAC, 5*(3), 216-226.

Steinwald, B., & Dummit, L. A. (1989). Hospital case-mix change: sicker patients or DRG creep? *Health Affairs, 8*(2), 35-47.

Stern, R. S., & Epstein, A. M. (1985). Institutional responses to prospective payment based on diagnosis-related groups: implications for cost, quality, and access. *Hospital Topics, 63*(3), 18-24.

Stortingsforhandling (1996). St.meld. nr. 44 (1995-96) Ventetidsgarantien - kriterier og finansiering [Norwegian Parliament Report No 44 Waiting time guarantee - Criterea and funding].

Stortingsforhandling (2015). Meld.St. 11 (2015-2016) Melding til Stortinget. Nasjonal helse- og sykehusplan (2016-2019) [National Health and Hospital Plan].

Street, A., O'Reilly, J., Ward, P., & Mason, A. (2011). DRG-based hospital payment and efficiency: theory, evidence, and challenges. *2011) Diagnosis-related groups in Europe: moving towards transparency, efficiency and quality in hospitals. Open University Press, Maidenhead*, 93-114.

Street, A., Vitikainen, K., Bjorvatn, A., & Hvenegaard, A. (2007). Introducing activity-based financing: a review of experience in Australia, Denmark, Norway and Sweden. The University of York. Centre for Health Economics. CHE Research Paper 30.

Tatchell, M. (1983). Measuring hospital output: a review of the service mix and case mix approaches. *Social science & medicine, 17*(13), 871-883.

Tjerbo, T., & Hagen, T. P. (2009). Deficits, soft budget constraints and bailouts: Budgeting after the Norwegian hospital reform. *Scandinavian Political Studies, 32*(3), 337-358.

Valdmanis, V. G., Rosko, M. D., & Mutter, R. L. (2008). Hospital quality, efficiency, and input slack differentials. *Health services research, 43*(5p2), 1830-1848.

Varabyova, Y., & Schreyögg, J. (2013). International comparisons of the technical efficiency of the hospital sector: panel data analysis of OECD countries using parametric and non-parametric approaches. *Health Policy, 112*(1), 70-79.

Vitikainen, K., Street, A., & Linna, M. (2009). Estimation of hospital efficiency—Do different definitions and casemix measures for hospital output affect the results? *Health Policy, 89*(2), 149-159.

Woolhandler, S., Ariely, D., & Himmelstein, D. U. (2012). Why pay for performance may be incompatible with quality improvement. *Bmj, 345*(7870), e5015.

WorldHealthOrganization, & StatensHelsetilsyn (1998). ICD-10: Den internasjonale statistiske klassifikasjon av sykdommer og beslektede helseproblemer. *The ICD-10: The international statistical classification of diseases and related health problems, Norwegian Board of Health Supervision). Oslo: Elanders forlag*.

Yin, J., Lurås, H., Hagen, T. P., & Dahl, F. A. (2013). The effect of activity-based financing on hospital length of stay for elderly patients suffering from heart diseases in Norway. *BMC health services research, 13*(1), 172.

10 Appendix

Paper I online supplementary material, available online 30 January 2017

http://www.sciencedirect.com/science/article/pii/S0168851017300246#appd002

Online supplementary material 1: Input and output data

This online supplementary material describes in more detail how data on the hospital activity were selected and regrouped for the productivity analyses.

Output data

Output data were provided by the Norwegian Patient Register. We selected 80,688,010 hospital episodes to be used in these analyses. All episodes in hospitals were grouped with a software program, a "DRG-grouper", which assigned each episode to a diagnostic-related group according to the age and sex of the patient, length of stay, diagnosis and procedures. In Norway, outpatients were only completely DRG grouped (and thus fully funded by the DRG system) from 2009 to 2014. To enable comparison across episode types and irrespective of the annual changes in financing, we regrouped all episodes using the same version of the grouper software for all years. This regrouping was done by the authors based upon classifications by the Nordic CaseMix Centre. Grouper version NOR2011co1F was supplied by DataWell Oy, an independent software firm based in Finland. Results from the regrouping included 761 DRGs with the same definition, which allowed for a stable comparison over time. A similar use of the same grouper was done by Kittelsen et al. [13].

Each of these groups had a weight corresponding to the relative cost of the treatment of the patients in that group. These weights determined the remuneration given to the hospitals and were adjusted annually (based on costs 2 years prior). Using these yearly weights of the original grouping will cause systematic errors because the annual changes in weights may be related to past years' productivity, technical change and time-specific financing. An even greater problem is that these weights do not handle the larger shifts taking place in the period, especially changes in the financing of day care treatments and outpatients. We calculated fixed weights based on average reimbursed DRG weights in 2011 (from the original grouping), and we used these weights for the regrouped data. Eleven minor groups had no cases in 2011 and, thus, we had no information about these weights. These groups were assigned weights from the years in which they occurred.

We classified DRG groups numbered 700–999 as outpatient episodes. Episodes in DRG groups numbered 1 – 699 were inpatients, aside from O-groups which were day care versions of regular inpatient DRGs. The four outputs thus measured the number of case-mix-adjusted treatments in each group. All cases in DRG409O (radiotherapy) were excluded because the operating costs for these treatments were not part of the inputs.

Healthy newborns were not registered in the output in 1999–2001 but were a part of the operating costs at the hospitals. To make the data comparable, we generated data on about 130,000

infants for the years 1999–2011 using public statistics on the number of births, and each hospital's share of births in the years for which we had data. The reimbursements for these newborns were part of the DRG weight of the mothers' DRG, which were higher in those years. From 2002, the healthy newborns (without complications) were registered in DRG 391 and were reimbursed normally in the DRG system, and the DRG weights of mothers and the newborns were recalibrated by lowering the weights of the mothers.

Hospital inputs

Hospital inputs usually include physical inputs such as human resources (number of staff in different categories), medical supplies/medicines, hospital beds and other categories of actual inputs. However, we measured inputs as operational costs to ensure comparability. Hospital costs are easily available, have the same data source all years and can thus be made comparable over a large timespan with better quality than other categories of physical inputs.

We defined operational costs as the total (accounted) expenses at the hospital, with reductions for teaching and research, non-treatment-related costs and capital costs. These costs were collected from annually published reports on Norwegian hospital care [19]. A similar approach for hospital inputs has been used by [13, 20, 21, 30]. Teaching and research costs were estimated by the grants given to hospitals for teaching and research but were not the actual teaching and research costs, which might have been higher and might have introduced a negative bias for large university hospitals. These hospitals may also have had a higher case mix. Overall, we made better adjustments in the earlier years than in later years because of the specific funding schemes and the way research grants were accounted for by hospital trusts. This may have introduced a negative trend bias for all hospitals doing research.

Further, to make the costs comparable over time, a deflator was applied to account for the annual increase in prices not related to changes in the level of output. For the years 2004–2014, the source of the deflator was Statistics Norway [19], and for the earlier years 1999–2003, the deflator was based on data included in reference [37]. All costs presented in this paper are real 2014 costs. Table A1 shows the percentage increase each year. The laws, regulations and accounting rules regarding pension costs changed considerably over these years, which caused systematic shifts for several years. These changes were captured in the deflator and, for the years 2010 and 2014, the pension costs have been estimated.

Table A1 Annual deflator, percentage increase from the preceding year

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
7.0	2.9	8.1	6.9	5.2	2.9	4.2	8.0	6.0	2.7	2.5	5.5	4.0	3.9	3.5

Online supplementary material 2: Further results

Table A2 Mean overall average input scale efficiency and estimated optimal hospital size (budget size)

Year	Sample average unit: input scale efficiency	Estimated optimal hospital size (MNOK)
1999	0.944 (0.917–0.968)	223 (182–517)
2000	0.913 (0.881–0.943)	189 (149–275)
2001	0.837 (0.807–0.866)	180 (141–449)
2002	0.899 (0.871–0.931)	204 (149–592)
2003	0.969 (0.947–0.998)	496 (321–992)
2004	0.932 (0.897–0.968)	589 (311–1,152)
2005	0.934 (0.903–0.966)	665 (354–1,255)
2006	0.937 (0.911–0.968)	724 (436–1,244)
2007	0.942 (0.918–0.969)	716 (398–1,255)
2008	0.973 (0.951–1.002)	753 (392–1,522)
2009	0.995 (0.957–1.050)	666 (330–1,560)
2010	1.002 (0.959–1.058)	673 (366–1,398)
2011	0.992 (0.960–1.046)	644 (399–1,249)
2012	0.994 (0.957–1.061)	606 (389–1,089)
2013	0.999 (0.968–1.055)	639 (401–1,155)
2014	1.000 (0.968–1.053)	629 (399–1,098)

Table notes: Optimal hospital size was calculated based on hospital-estimated efficiency multiplied by yearly costs divided by relative distance from scale efficiency. The confidence interval for estimated hospital size was calculated from bootstrapping of the relative distance from scale efficiency.

Paper I

Health Policy 121 (2017) 418–425

Contents lists available at ScienceDirect

Health Policy

journal homepage: www.elsevier.com/locate/healthpol

Productivity growth, case mix and optimal size of hospitals. A 16-year study of the Norwegian hospital sector

Kjartan Sarheim Anthun[a,b,*], Sverre Andreas Campbell Kittelsen[c], Jon Magnussen[a]

[a] Norwegian University of Science and Technology, Department of Public Health and Nursing, PO Box 8905, 7491 Trondheim, Norway
[b] SINTEF Technology and Society, Department of Health Research, PO Box 4760, Sluppen, 7465 Trondheim, Norway
[c] The Ragnar Frisch Centre for Economic Research, Gaustadalléen 21, 0349 Oslo, Norway

ARTICLE INFO

Article history:
Received 9 November 2016
Received in revised form 17 January 2017
Accepted 19 January 2017

Keywords:
Health care reform
Health services research
Organizational efficiency
Healthcare financing
Diagnosis-related groups
Hospital economics

ABSTRACT

Background and objectives: This paper analyses productivity growth in the Norwegian hospital sector over a period of 16 years, 1999–2014. This period was characterized by a large ownership reform with subsequent hospital reorganizations and mergers. We describe how technological change, technical productivity, scale efficiency and the estimated optimal size of hospitals have evolved during this period.
Material and methods: Hospital admissions were grouped into diagnosis-related groups using a fixed-grouper logic. Four composite outputs were defined and inputs were measured as operating costs. Productivity and efficiency were estimated with bootstrapped data envelopment analyses.
Results: Mean productivity increased by 24.6% points from 1999 to 2014, an average annual change of 1.5%. There was a substantial growth in productivity and hospital size following the ownership reform. After the reform (2003–2014), average annual growth was <0.5%. There was no evidence of technical change. Estimated optimal size was smaller than the actual size of most hospitals, yet scale efficiency was high even after hospital mergers. However, the later hospital mergers have not been followed by similar productivity growth as around time of the reform.
Conclusions: This study addresses the issues of both cross-sectional and longitudinal comparability of case mix between hospitals, and thus provides a framework for future studies. The study adds to the discussion on optimal hospital size.

1. Introduction

In the past 20 years, many countries have undergone large changes in the way health care is organized, financed and delivered. Under the umbrella of new public management, there has been an increase in quasi-markets, choice and competition, and increased use of activity- and results-based financing. In traditionally public tax-based systems, such as the UK and Norway, public hospitals have been reorganized into trusts with a large degree of autonomy [1,2]. At the same time, several countries have pursued a policy of centralization, both in terms of exploiting perceived scale efficiency in the provision of services and by

* Corresponding author at: Norwegian University of Science and Technology, Department of Public Health and Nursing, PO Box 8905, Trondheim 7491, Norway.
E-mail addresses: kjartan.s.anthun@sintef.no, kjartan.anthun@ntnu.no (K.S. Anthun), sverre.kittelsen@frisch.uio.no (S.A.C. Kittelsen), jon.magnussen@ntnu.no (J. Magnussen).

http://dx.doi.org/10.1016/j.healthpol.2017.01.006
0168-8510/

shifting power from local to central authorities [2,3]. The recession in 2009 spurred a policy of fiscal austerity that has put health care, together with other publicly funded welfare services, under pressure.

Health care reforms, as well as increased fiscal pressure, infer an increased focus on how resources allocated to health care are used. Efficient use of available resources is an important policy goal in all health care systems. Regulators and policy makers will typically be interested in the level of productivity, whether and at what rate productivity increases or decrease over time, and the relationships between productivity and different regulatory, structural and financial policy measures. Hospitals constitute a major part of the health care sector; therefore, policy makers are particularly interested in assessing their performance. However, comparisons of productivity across hospitals are inherently difficult because of differences in case mix. Differences in case mix are often controlled for by using patient classification systems such as the diagnosis-related groups (DRGs) when describing hospital activity [4].

In this paper, we describe and discuss how hospital productivity has evolved in Norway from 1999 to 2014. Our analysis of this 16-year period enabled us to look at productivity from the long-term perspective of a period that included one major health care reform. Before 2002, all hospitals were owned and operated by the counties, and the hospitals had long waiting times and large deficits. The counties could not levy taxes themselves, so there was gaming of the budgeting and consequently soft budgeting as additional funding was provided by the central government [3,5,6]. In 2002, hospital ownership was transferred from 19 counties to the central government, and the responsibility for the provision of services was given to five (currently four) regional health authorities. The regional health authorities organized hospitals through hospital trusts.

Following the reform, there was major structural changes as the number of hospital trusts has decreased through mergers and reorganizations. Some minor hospitals that were located near larger hospitals were closed after mergers and reorganization of services. As a result, several hospital trusts are now multi-sited hospitals, and some even administer several multi-sited hospitals. According to Jacobs et al. [7], entities used in productivity analyses must have discretion about the conversion from inputs to outputs, must capture the entire production process and must be comparable. This applies to hospitals, hospital trusts and multi-site hospital trusts. Throughout this paper, we denote the organizational units as "hospitals" while acknowledging that these units often encompass several locations or physical hospitals.

This paper comprises three parts. First, we propose a way of describing hospital activity that captures both longitudinal and cross-sectional differences in case mix. This is crucial for capturing the effects of changes in treatment procedures on hospital productivity and enables us to relate the observed changes in productivity to the institutional and structural changes that have taken place during this period. In addition to the hospital reform in 2002, there has also been a substantial transition from inpatient to day care and outpatient treatment. To determine the long-term effects of reforms and policy changes, it is important to use data over a long time span. This is not commonly done, and most hospital efficiency analyses are either cross-sectional or span 1 year before and 1 year after a reform [8,9]. In this analysis, we used data envelopment analysis (DEA) with a long time series and case mix-adjusted output measures. A long term approach was also presented by Halsteinli et al. [10], who used data for 9 years in their analysis of child and adolescent mental health services, and that of Biørn et al. [11], who used a 10-year span when evaluating a hospital financing reform. However, these two Norwegian studies did not adequately adjust for potential longitudinal changes in case mix.

Second, we estimate Malmquist indices [12,13] to analyse to what extent the observed changes in productivity resulted from technical change (the best becoming better) or from changes in relative efficiency ("catching up"). Technical change is based on the performance of the best practice hospitals and it is as such it is often the result of a general development rather than the institutional environment or local policy initiatives. Thus, the relative share of the catching up and technical change elements provide an indication of the relationships between institutional environment, policy measures and provider performance.

Third, following the reform in 2002, the average hospital size has increased substantially, through reorganizations, mergers and hospital closures. There are arguments both for economies and diseconomies of scale in the literature [13–15], and we measured scale efficiency and estimated optimal scale and tracked the changes in these variables.

It is difficult to hypothesize the effects of the reform on catching up or technical change because the reformed implied both centralization and decentralization. If the governance of hospitals is strengthened, we might expect increased homogeneity in the results and thus reduced variance behind the frontier. However, technical change may not necessarily coincide with the reform if efficient hospitals already have exhausted their potential for improvement.

2. Materials and methods

2.1. Measuring hospital inputs and outputs

There are two issues that must be dealt with in an analysis of productivity growth. First, differences in case mix *between hospitals* must be adjusted for. Relative cost-weights (DRG prices) can be used to aggregate individual episodes into larger groups of hospital activity. However, such aggregation requires the assumption that relative treatment costs are independent of hospital case mix and size. Moreover, the results are usually sensitive to the type of case-mix adjustment that is chosen [8,16–18]. Too many output categories can artificially inflate the number of efficient hospitals because rarer combinations of outputs determine the estimated best-practice front. Using all DRG groups as output dimensions would give no degrees of freedom because the number of DRGs would surpass the number of hospitals. A different approach would be to aggregate all hospital activity into one group, but that would underestimate differences between hospitals. Our

point of departure is that the aim of any aggregation of groups should be to capture essential structural differences between hospitals. The distinguishing trends in hospital care during this period were the decreased length of stay and increased use of same-day surgery and outpatient treatments. The composite types listed below are well suited for capturing these trends, while at the same time allowing hospitals with different strategies related to the types of treatment to be considered efficient. We distinguish here between four composite types of hospital activity to summarize the activity and the technological profile of each hospital:

- Emergency inpatient discharges.
- Elective inpatient discharges.
- Day care treatments.
- Outpatient visits and treatments.

The second issue is related to case-mix comparability *between years.* Over the period covered in this analysis, there have been changes in the documentation of diagnoses and procedures, changes in DRG rules (grouper logic) and technological changes (which may shift patients between different types of hospital activities). The funding scheme was set up to accommodate this by annual adjustments of relative prices and, when needed, by changing the DRG logic (i.e., which procedures and treatments belong to which group). Consequently, technological changes may be captured by the annual changes in either DRG logic or the cost weights, and we would be unable to capture their effect on productivity. To avoid this, we regrouped all hospital episodes using a common grouper. We also applied fixed cost-weights. Technological change was thus captured as productivity growth and was not "hidden" in changes in relative prices or changes in grouper logic. For more details, we refer to the online Supplementary material which describes in detail the output and input data.

Inputs were measured as deflated total hospital operating costs. These adjusted costs only included production in the four composite outputs; i.e., only DRG activity [19–21]. These costs are routinely used in productivity analyses and have been shown to have good comparability across time and units [19–21]. Capital costs and personnel were not used as inputs because of the lack of longitudinal comparability as data were available only for shorter time spans. The source of the hospital costs was Statistics Norway.

Information about hospital episodes in all Norwegian somatic hospitals in the period 1999–2014 were provided by the Norwegian Patient Register. Hospitals providing only elective care were excluded.

We aggregated the hospital activity data to the level of the cost data. The level of analysis was the operational ownership level; i.e., hospitals before the hospital ownership reform in 2002 and hospital trusts after the reform. There were 506 observations after aggregation.

2.2. Methods: estimating productivity and efficiency

Productivity is the relationship between inputs and outputs, whereas efficiency is the relationship between the observed productivity and the best possible productivity. A production frontier is the boundary of the production possibility set; the frontier thus describes the optimal possibility of conversion of inputs to outputs. Two related but distinctively different methodologies are commonly used for estimating frontiers and subsequently efficiency [7,22–24]. Stochastic frontier analysis (SFA) is a parametric approach in which the efficiency frontier has a specific functional form, and the method incorporates random errors in the estimation of the production function. DEA is the other main approach. First suggested in 1957 [25], DEA was developed in 1978 [26]. The method has been used at different levels for comparing nations [27] and regions [28] but more commonly hospitals [23]. In Norway and the Nordic countries, various studies have used DEA for estimating hospital productivity [11,13,16,21,29–33]. This method is favoured in environments in which substantial measurement errors are unlikely and with multiple inputs and/or outputs [9,24,34]. This is assumed to apply to the public Norwegian hospital sector with the data described above, and we have thus chosen the DEA approach.

We estimated a production frontier and compared each hospital's annual production to that frontier. A hospital on the frontier is, by definition, efficient and has a DEA score of 1.0. A hospital behind the frontier is considered inefficient and has a score <1.0, which represents the inefficiency as calculated as the relative distance to the frontier. As an example, a DEA score of 0.9 is interpreted as 10% inefficiency.

The Malmquist index of productivity growth compares the same decision-making unit between two years t and t+1. This index is then decomposed into an efficiency change and technical change. In the analysis of change, we compared two years so that a result >1.0 indicates increased productivity and <1.0 indicates decreased productivity. We assumed sequential (accumulated) frontiers. The productivity of a unit may be compared with best-practice hospitals at the frontier in the same year and in earlier years but never with observations in the future. It is likely that the productivity of a hospital in 1999 may be attainable by hospitals in 2014, but the opposite will generally not be the case.

In economic terms, it is common to distinguish between constant returns to scale (CRS) and variable returns to scale (VRS). CRS imply a linear conversion from inputs to outputs, whereas VRS technology allows a differentiated conversion ratio depending on the size of the hospital. For instance, fixed costs often imply increasing economies of scale where hospitals can increase their productivity by producing more as average costs decrease with increased volume. The term VRS is the general case, and CRS is a testable special case. In our analysis, we tested and failed to reject VRS, and therefore measured *scale efficiency* as the distance between the VRS and CRS frontier.

We estimated the optimal size of a hospital by multiplying each hospital's observed productivity by its cost, and then dividing by the relative distance to the optimal scale-efficient hospital. A confidence interval was created based on the distribution of the bootstrapped estimate of the relative distance to the optimal scale-efficient hospital.

When summarizing results from the analyses we calculate the annual mean level of productivity weighted by

Table 1
Output shares, relative hospital size and number of hospitals, N = 506.

Year	Emergency	Elective	Day care	Outpatient	Size (1999 = 1)	#Hospitals
1999	0.51	0.32	0.05	0.12	1.0	55
2000	0.50	0.32	0.05	0.13	1.0	54
2001	0.49	0.33	0.05	0.13	1.1	54
2002	0.49	0.33	0.05	0.12	1.7	36
2003	0.47	0.34	0.06	0.12	2.1	30
2004	0.48	0.34	0.06	0.13	2.1	30
2005	0.48	0.33	0.06	0.13	2.2	29
2006	0.48	0.33	0.06	0.13	2.3	29
2007	0.48	0.32	0.06	0.14	2.4	28
2008	0.48	0.32	0.06	0.14	2.4	28
2009	0.48	0.30	0.06	0.15	3.1	23
2010	0.49	0.29	0.07	0.15	3.3	22
2011	0.47	0.30	0.07	0.17	3.3	22
2012	0.46	0.30	0.07	0.17	3.5	22
2013	0.46	0.30	0.07	0.17	3.3	22
2014	0.45	0.30	0.07	0.18	3.4	22

Note: hospital size was measured as mean real operating costs relative to the 1999 mean.

hospital operating cost. However, the annual scale efficiency of the average hospital is based on the sample average unit (SAU, which provides an implicit weighting). To calculate confidence intervals, the sample results are bootstrapped [35,36].

3. Results

Table 1 shows the shares of each output relative to total case-mix adjusted number of episodes, and how average hospital size has changed during the period of analysis. The average hospital has more than tripled in size since 1999, primarily through reorganizations and mergers as the number of hospitals have more than halved in this period from 55 in 1999 to 22 in 2014. Outpatient treatment has increased in importance as the relative volume has increased by 45% since 1999. The share of day care of the total production gained 28%, while the share of inpatient treatment has reduced by 11% for emergencies and 5% for elective treatments.

In Fig. 1 the development in productivity is shown with 1999 normalized to 1. We find an overall increase in annual mean productivity level by 24.6% points from 1999 to 2014. There is a wide confidence interval, thus not all annual changes are significantly different from the previous year. In the years prior to the hospital reform, there are no significant changes in productivity. However, we find a substantial shift from 2002 to 2003, and also a positive trend from 2004 until 2014.

Measures of total productivity growth, technical change and catching up are presented in Table 2. We found a large improvement in productivity around the time of the reform, with annual changes of 5–6% from 2001 to 2002 and from 2002 to 2003. However, in the years after 2003, annual changes were generally small, and in the whole period three of the years had a negative productivity growth (2000/2001, 2008/2009 and 2011/2012). The estimates for annual front shifts were significant only for 1999/2000 and 2003/2004. The estimates (and confidence intervals) for efficiency change for five of the years was >1.0, indicating that the mean hospital moved closer to the frontier. In these years, the hospitals were "catching up". In two years, the efficiency change was <1.0, which indicated that the mean hospital was not catching up but was "falling behind".

Although the mean size of hospitals more than tripled during this period, the average scale efficiency remained stable. There was a slightly higher scale efficiency in the period 2008–2014 than in 1999–2007. In 1999, 10 of the 55 hospitals were smaller than the estimated optimal hospital size, but in 2003 only three of the 30 hospitals were smaller than the estimated optimal hospital size. In 2012–2014, all of the hospitals were larger than the estimated optimal hospital size. Fig. 2 shows the actual observed hospital sizes and the estimated optimal size (including 95% confidence intervals for the estimated optimal size). More detailed numbers relating to the scale efficiency and optimal size are available as online Supplementary material.

4. Discussion

The main policy reform relevant to this analysis was the hospital reform in 2002. The effects of the reform on hospital productivity were not obvious a priori, but one of the stated goals was to improve efficiency. It has been argued that the reform included both elements of centralization (state ownership) and decentralization (regional health authorities) [3]. Through centralization of the purchaser and provider role, productivity was expected to increase [37]. Another goal of the reform was to replace soft budgeting with hard budgeting. Early empirical work indicated that productivity increased by about 5% after the reform [37]. This present study confirms that the major shift in productivity in the 16-year period covered here occurred around the time of the reform. Overall, we found a productivity increase of 24.6% points from 1999 to 2014. In the years after the reform (2003–2014), we found an average annual growth of <0.5%.

Separating productivity growth into "catching up" (the less efficient hospitals improving) and "technical change" (the production frontier shifting outwards) may give important information for policy makers. However, the results presented here fail to provide a clear picture. Mostly

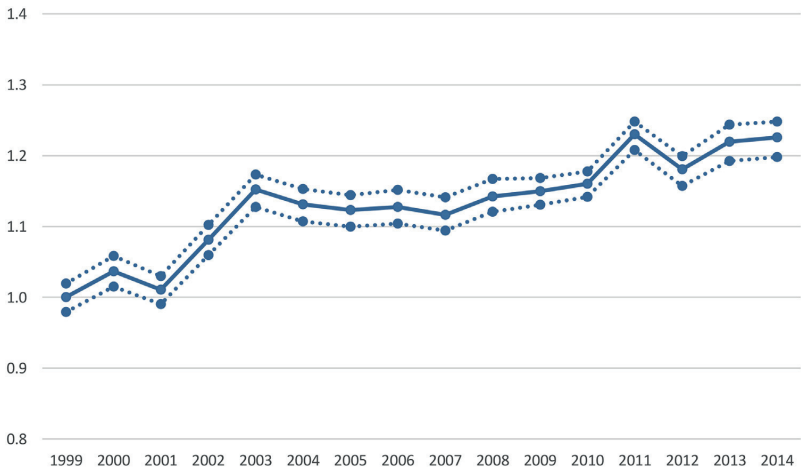


Fig. 1. Bootstrapped productivity estimates, weighted mean unit, pooled base, and 95% confidence intervals, relative development from 1999.

Table 2
Decomposition of productivity growth (M) into catching up (MC) and front shift (MF), with 95% confidence intervals and weighted mean.

Year	N	Malmquist index of productivity growth	Efficiency change/catching up	Technical change/front shift
1999/2000	54	1.046 (1.034–1.059)	0.963 (0.920–1.014)	1.086 (1.026–1.133)
2000/2001	53	0.971 (0.960–0.982)	0.966 (0.944–1.004)	1.005 (0.962–1.031)
2001/2002	19	1.056 (1.045–1.067)	1.069 (1.048–1.109)	0.987 (0.947–1.011)
2002/2003	30	1.062 (1.047–1.077)	1.024 (0.992–1.066)	1.038 (0.992–1.074)
2003/2004	30	1.000 (0.988–1.011)	0.936 (0.896–0.985)	1.069 (1.013–1.113)
2004/2005	28	1.007 (0.996–1.017)	1.018 (0.990–1.062)	0.989 (0.943–1.016)
2005/2006	29	1.008 (0.993–1.021)	1.003 (0.976–1.048)	1.005 (0.961–1.028)
2006/2007	28	1.005 (0.989–1.021)	1.009 (0.985–1.049)	0.996 (0.959–1.017)
2007/2008	28	1.021 (1.011–1.030)	1.028 (1.011–1.063)	0.993 (0.960–1.007)
2008/2009	21	0.981 (0.966–0.995)	0.984 (0.966–1.007)	0.996 (0.974–1.015)
2009/2010	21	1.001 (0.990–1.012)	0.988 (0.972–1.016)	1.013 (0.987–1.030)
2010/2011	22	1.054 (1.036–1.069)	1.032 (1.004–1.071)	1.021 (0.989–1.047)
2011/2012	22	0.958 (0.948–0.968)	0.963 (0.947–0.994)	0.995 (0.966–1.008)
2012/2013	22	1.037 (1.025–1.048)	1.042 (1.027–1.074)	0.995 (0.967–1.007)
2013/2014	22	1.006 (0.997–1.016)	0.996 (0.982–1.027)	1.010 (0.984–1.016)

Note: the numbers represent a comparison of two consecutive years for each hospital, >1.0 indicates productivity growth and <1.0 indicates productivity decline.

the estimates were not statistically significant, and for those that were, it was difficult to see a clear pattern and to link this to plausible policy explanations.

Following the hospital reform in 2002, the number of hospitals decreased substantially. This reduction was mainly the result of organizational mergers, although these were also accompanied by internal restructuring. The number of hospitals more than halved as the mean hospital size more than tripled. A major motivation for the transfer of hospital ownership to the state was to avoid unnecessary duplication of services across hospitals. Thus, we would expect scale efficiency to have increased and consequently that actual hospital size would have moved closer to the optimal hospital size. Although somewhat higher in the last years of our study, scale efficiency did not change significantly during the period. This suggests that merging hospitals is not a recipe for achieving efficiency. In our study, the aggregated productivity growth around the reform coincided with mergers, but later mergers did not coincide with similar growth.

There was larger activity growth in 2001–2003 than what was planned [5] due to budget gaming and soft budgeting constraint. Although this is also true for the costs (and subsequent deficits), hospitals seemed to increase size and productivity around the time of the reform. This could indicate that hospitals exploited economies of scale.

In this paper, the estimated optimal hospital size was quite small and, for the years 2012–2014, all of the hospitals were larger than the estimated optimal hospital size. However, the average optimal size did increase as the number of hospitals decreased from 2002 to 2005. This reflects that some of the merged hospitals performed well enough to define the optimal size from that point in time. However, these estimates rely on good case-mix adjustments, particularly for the largest hospitals. If cases were more severe or quality higher *within each DRG* in large than in small

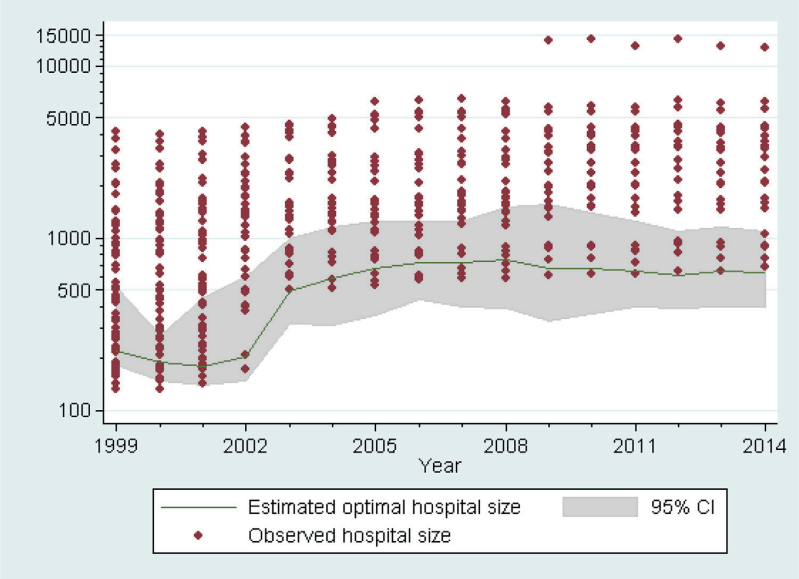


Fig. 2. Observed hospital size and estimated optimal hospital size with 95% confidence intervals. The data are expressed on a logarithmic scale for hospital size, as measured by real operating costs (million NOK).

hospitals, the data do not fully capture the production of the large hospitals, and consequently, our estimate of optimal scale would have been downwards biased. However, a recent study reported no association between hospital size and differences in productivity, and no clear cost-quality trade-off pattern [30].

We have measured productivity at the operational level to compare hospitals before the 2002 reform with hospital trusts after the reform. Depending on the distribution behind the frontier, the level of analysis could cause some differences in the level of productivity; however, the development will be the same. The frontier depends only on the best units and is not as sensitive to marginal changes in the number of observations as are other methods. The scale efficiency measures should not depend on the level of analysis. Because the productivity change (as measured by the Malmquist index) is a product of scale change, technical change and pure efficiency change, it is possible to increase the overall productivity without an increase in all of the three components, and thus the productivity seemed to increase despite the fact that hospitals were larger than the estimated optimal size. Overall, these results for scale and size are consistent with the report by Kittelsen et al. [13], who estimated scale efficiency in Norway as comparatively high and found that hospitals were larger than the estimated optimal size.

Changes in the diagnostic registrations might be one explanation for the observed productivity growth. Norwegian hospitals are funded by a combination of fixed budgets and activity-based financing, and thus income relies partly on the recording of diagnoses and procedures. Recording of more diagnoses and procedures may have led to an increase in outputs through more expensive DRGs and subsequently higher estimates of technical productivity. However, this effect would only be short term as the prices are updated after 2 years to reflect average costs for each treatment in each group. Any large-scale wrongful upcoding would thus not yield long-term effects. There is currently more use of secondary diagnoses than in the early years, but this shift is linked only to a small extent to DRG-level prices [38]. In the present study, we used a common grouper and fixed weight for all years to avoid some of the issues related to upcoding. However, even after regrouping the data, we cannot exclude the possibility that some changes occurred because of upcoding rather than real changes in activity.

Some recent studies have incorporated quality indices as an output measure or have otherwise controlled for quality [30,39,40]. In this paper, we have measured only the hospital production by volume based on average costs, not the contents of that production. Both quality and actual health improvements are very important and may have changed over this long time span, and we suggest that further research is needed to answer this question.

In this paper, we have improved the comparability of the output measurement by using a fixed DRG logic and weights for all years. This enabled us to use a dataset of 16 years for hospital-level analyses which, to our knowledge, is an unprecedented long time span for hospital-level productivity analyses. We believe that it is important in studies of reform to apply a longitudinal perspective instead of analysing only 1 year before and 1 year after the reform.

5. Conclusion

The present study shows that mean productivity increased by 24.6% points from 1999 to 2014, an average annual change of 1.5%. There was a substantial growth following the ownership reform in 2002. After the reform (2003–2014), the average annual growth was <0.5%. There was no evidence of technical change. The estimated optimal size is smaller than the actual size of most hospitals.

Conflict of interest

The authors report no conflicts of interest.

Acknowledgements

This work was supported by The Research Council of Norway, grant number 214338. The authors wishes to thank Lise Rochaix and Youngho Oh who commented valuably on earlier drafts of this paper. Valuable comments were also provided by two anonymous reviewers of this journal.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.healthpol.2017.01.006.

References

[1] Martinussen P, Magnussen J. Health Care Reforms: The Nordic Experience. Nordic Health Care Systems Recent Reforms and Current Policy Challenges European Observatory on Health Systems and Policies Series. Maidenshead, Berkshire: Open University Press—McGraw-Hill Education; 2009. p. 21–52.
[2] Saltman RB, Bankauskaite V, Vrangbæk Ke. Decentralization in Health Care—Strategies and Outcomes. Mc Graw Hill Open University Press; 2007.
[3] Magnussen J, Hagen TP, Kaarboe OM. Centralized or decentralized? A case study of Norwegian hospital reform. Social Science & Medicine 2007;64:2129–37.
[4] Geissler A, Quentin W, Scheller-Kreinsen D, Busse R. Introduction to DRGs in Europe: common objectives across different hospital systems. In: Busse R, Geissler A, Quentin W, Wiley M, editors. Diagnosis Related Groups in Europe: Moving Towards Transparency, Efficiency and Quality in Hospitals. 2011. p. 9–21.
[5] Tjerbo T, Hagen TP. Deficits soft budget constraints and bailouts: budgeting after the Norwegian hospital reform. Scandinavian Political Studies 2009;32:337–58.
[6] Hagen TP, Kaarbøe OM. The Norwegian hospital reform of 2002: central government takes over ownership of public hospitals. Health Policy 2006;76:320–33.
[7] Jacobs R, Smith PC, Street A. Measuring Efficiency in Health Care: Analytic Techniques and Health Policy. Cambridge University Press; 2006.
[8] Chowdhury H, Zelenyuk V, Laporte A, Wodchis WP. Analysis of productivity, efficiency and technological changes in hospital services in Ontario: how does case-mix matter. International Journal of Production Economics 2014;150:74–82.
[9] O'Neill L, Rauner M, Heidenberger K, Kraus M. A cross-national comparison and taxonomy of DEA-based hospital efficiency studies. Socio-Economic Planning Sciences 2008;42:158–89.
[10] Halsteinli V, Kittelsen SA, Magnussen J. Productivity growth in outpatient child and adolescent mental health services: the impact of case-mix adjustment. Social Science & Medicine 2010;70:439–46.
[11] Biørn E, Hagen TP, Iversen T, Magnussen J. The effect of activity-based financing on hospital efficiency: a panel data analysis of DEA efficiency scores 1992–2000. Health Care Management Science 2003;6:271–83.
[12] Färe R, Grosskopf S, Lindgren B, Roos P. Productivity Changes in Swedish Pharamacies 1980–1989: A Non-Parametric Malmquist Approach. International Applications of Productivity and Efficiency Analysis. Springer; 1992. p. 81–97.
[13] Kittelsen SA, Winsnes BA, Anthun KS, Goude F, Hope Ø, Häkkinen U, et al. Decomposing the productivity differences between hospitals in the Nordic countries. Journal of Productivity Analysis 2015;43:281–93.
[14] Banker RD. Estimating most productive scale size using data envelopment analysis. European Journal of Operational Research 1984;17:35–44.
[15] Dranove D. Economies of scale in non-revenue producing cost centers: implications for hospital mergers. Journal of Health Economics 1998;17:69–83.
[16] Magnussen J. Efficiency measurement and the operationalization of hospital production. Health Services Research 1996;31:21.
[17] Rosen AK, Loveland SA, Rakovski CC, Christiansen CL, Berlowitz DR. Do different case-mix measures affect assessments of provider efficiency? Lessons from the department of veterans affairs. The Journal of Ambulatory Care Management 2003;26:229–42.
[18] Vitikainen K, Street A, Linna M. Estimation of hospital efficiency—do different definitions and casemix measures for hospital output affect the results. Health Policy 2009;89:149–59.
[19] Samdata. SAMDATA Spesialisthelsetjenesten [Norwegian annual statistics on specialized hospital care]. Norwegian Directorate of Health, IS-2348; 2015.
[20] Anthun KS, Goude F, Häkkinen U, Kittelsen S, Kruse M, Medin E, et al. Eurohope Hospital Level Analysis: Material, Methods and Indicators. Eurohope Discussion Papers Helsinki. THL; 2013.
[21] Medin E, Häkkinen U, Linna M, Anthun KS, Kittelsen SA, Rehnberg C, et al. International hospital productivity comparison: experiences from the Nordic countries. Health Policy 2013;112:80–7.
[22] Fried HO, Lovell Cak, Schmidt SSe. The Measurement of Productive Efficiency and Productivity Growth. Oxford University Press; 2008.
[23] Hollingsworth B. The measurement of efficiency and productivity of health care delivery. Health Economics 2008;17:1107–28.
[24] Mortimer D. Competing Methods for Efficiency Measurement: A Systematic Review of Direct DEA vs SFA/DFA Comparisons; 2002.
[25] Farrell MJ. The measurement of productive efficiency. Journal of the Royal Statistical Society Series A (General) 1957;120:253–90.
[26] Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. European Journal of Operational Research 1978;2:429–44.
[27] Varabyova Y, Schreyögg J. International comparisons of the technical efficiency of the hospital sector: panel data analysis of OECD countries using parametric and non-parametric approaches. Health Policy 2013;112:70–9.
[28] Halkos GE, Tzeremes NG. A conditional nonparametric analysis for measuring the efficiency of regional public healthcare delivery: an application to Greek prefectures. Health Policy 2011;103:73–82.
[29] Biørn E, Hagen TP, Iversen T, Magnussen J. How different are hospitals' responses to a financial reform? The impact on efficiency of activity-based financing. Health Care Management Science 2010;13:1–16.
[30] Kittelsen SA, Anthun KS, Goude F, Huitfeldt I, Häkkinen U, Kruse M, et al. Costs and quality at the hospital level in the Nordic countries. Health Economics 2015;24:140–63.
[31] Linna M. Measuring hospital cost efficiency with panel data models. Health Economics 1998;7:415–27.
[32] Linna M, Häkkinen U, Peltola M, Magnussen J, Anthun KS, Kittelsen S, et al. Measuring cost efficiency in the Nordic Hospitals—a cross-sectional comparison of public hospitals in 2002. Health Care Management Science 2010;13:346–57.
[33] Medin E, Anthun KS, Häkkinen U, Kittelsen SA, Linna M, Magnussen J, et al. Cost efficiency of university hospitals in the Nordic countries: a cross-country analysis. The European Journal of Health Economics 2011;12:509–19.
[34] Hollingsworth B, Dawson P, Maniadakis N. Efficiency measurement of health care: a review of non-parametric methods and applications. Health Care Management Science 1999;2:161–72.
[35] Simar L, Wilson PW. Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. Management Science 1998;44:49–61.
[36] Simar L, Wilson PW. Statistical inference in nonparametric frontier models: the state of the art. Journal of Productivity Analysis 2000;13:49–78.
[37] Kittelsen SA, Magnussen J, Anthun KS, Häkkinen U, Linna M, Medin E, et al. Hospital Productivity and the Norwegian Ownership Reform: A Nordic Comparative Study. Stakes; 2008.

[38] Anthun KS, Bjørngaard JH, Magnussen J. Economic incentives and diagnostic coding in a public health care system. International Journal of Health Economics and Management 2016:1–19.
[39] Du J, Wang J, Chen Y, Chou S-Y, Zhu J. Incorporating health outcomes in Pennsylvania hospital efficiency: an additive super-efficiency DEA approach. Annals of Operations Research 2014;221:161–72.
[40] Ferrier GD, Trivitt JS. Incorporating quality into the measurement of hospital efficiency: a double DEA approach. Journal of Productivity Analysis 2013;40:337–55.

Paper II

J Prod Anal (2015) 43:281–293
DOI 10.1007/s11123-015-0437-z

Decomposing the productivity differences between hospitals in the Nordic countries

Sverre A. C. Kittelsen · Benny Adam Winsnes · Kjartan S. Anthun · Fanny Goude · Øyvind Hope · Unto Häkkinen · Birgitte Kalseth · Jannie Kilsmark · Emma Medin · Clas Rehnberg · Hanna Rättö

Published online: 26 February 2015

Abstract Previous studies indicate that Finnish hospitals have significantly higher productivity than in the other Nordic countries. Since there is no natural pairing of observations between countries we estimate productivity levels rather than a Malmquist index of productivity differences, using a pooled set of all observations as reference. We decompose the productivity levels into technical efficiency, scale efficiency and country specific possibility sets (technical frontiers). Data have been collected on operating costs and patient discharges in each diagnosis related group for all hospitals in the four major Nordic countries, Denmark, Finland, Norway and Sweden. We find that there are small differences in scale and technical efficiency between countries, but large differences in production possibilities (frontier position). The country-specific Finnish frontier is the main source of the Finnish productivity advantage. There is no statistically significant association between efficiency and status as a university or capital city hospital. The results are robust to the choice of bootstrapped data envelopment analysis or stochastic frontier analysis as frontier estimation methodology.

Keywords Productivity · Hospitals · Efficiency · DEA · SFA

JEL Classification C14 · I12

S. A. C. Kittelsen (✉) · B. A. Winsnes
Frisch Centre, Oslo, Norway
e-mail: sverre.kittelsen@frisch.uio.no

B. A. Winsnes
Oslo University Hospital, Oslo, Norway

K. S. Anthun · Ø. Hope · B. Kalseth
SINTEF, Trondheim, Norway

F. Goude · E. Medin · C. Rehnberg
Karolinska Institutet, Stockholm, Sweden

U. Häkkinen · H. Rättö
National Institute for Health and Welfare – THL, Helsinki, Finland

J. Kilsmark
Danish Institute of Health Research, Copenhagen, Denmark

1 Introduction

In previous studies (Kittelsen et al. 2009; Kittelsen et al. 2008; Linna et al. 2006, 2010) one has found persistent evidence that the somatic hospitals in Finland have a significantly higher average productivity level than hospitals in the other major Nordic countries (Sweden, Denmark and Norway).[1] These results indicate that there could be significant gains from learning from the Finnish example, especially in the other Nordic countries, but potentially also in other similar countries. The policy implications could however be very different depending on the source of the productivity differences. This paper extends earlier work by, (1) decomposing the productivity differences into those that stem from technical efficiency, scale efficiency and differences in the possibility set (the technology) between periods and countries, and (2) exploring the statistical associations between the technical efficiency and various hospital-level indicators such as case-mix, outpatient share and status as a university or capital city hospital.

[1] Although the Nordic countries also include Iceland, comparable data on Icelandic hospitals have not been available. In this article we will therefore use the term Nordic countries about the four largest countries.

Finally, (3) we examine the robustness of the results to the choice of method.

International comparisons of productivity and efficiency of hospitals are few, primarily because of the difficulty of getting comparable data on output (Derveaux et al. 2004; Linna et al. 2006; Medin et al. 2013; Mobley and Magnussen 1998; Steinmann et al. 2004; Varabyova and Schreyögg 2013). Such analyses often find quite substantial differences in performance between countries. Differences may be due to the dissimilar hospital structures and financing schemes, e.g. whether hospitals exploit economies of scale, have an optimal level of specialisation, or face high-powered incentive schemes that would encourage efficient production. Differences may also result from methodological problems. Cross-national analyses are often based on data sets that only to a limited extent are comparable—in the sense that inputs and outputs are defined and measured differently across countries. Our comparison gains validity from the existence of a Nordic standard for diagnosis related groups (DRGs) (Linna and Virtanen 2011). As described in the data section, the structure of the hospital sectors are broadly similar in the Nordic countries and the main differences are handled by assuming country specific production frontiers and variables in the analysis. It is, however, well known that the way we measure hospital performance may influence the empirical efficiency measures (Halsteinli et al. 2010; Magnussen 1996). In this article we will therefore use both the non-parametric data envelopment analysis (DEA) method and the stochastic frontier analysis (SFA) method, and provide evidence of the robustness of our results.

2 Methods

2.1 Efficiency and productivity

Efficiency and productivity are often used interchangeably. In our terminology productivity denotes the ratio of inputs and outputs, while efficiency is a relative measure comparing actual to optimal productivity. Since *productivity* is a ratio, it is by definition a concept that is homogenous of degree zero in inputs and outputs, i.e. a constant returns to scale (CRS) concept. This does not imply that the underlying technology is CRS. Indeed, the technology may well exhibit variable returns to scale (VRS), and equally efficient units may well have different productivity depending on their scale of operation, as well as other differences in their production possibility sets.

Most productivity indexes rely on prices to weigh several inputs and/or outputs, but building on Malmquist (1953), Caves et al. (1982) recognised that (lacking prices) one can instead use properties of the production function, i.e. rates of transformation and substitution along the frontier of the production possibility set, for an implicit weighting of inputs and outputs. We will use the term *technical productivity* to denote such a ratio of inputs to outputs where the weights are not input and output prices but rather derived from the estimated technologies.

This analysis departs from Farrell (1957) who defined (the input-oriented) *technical efficiency* as:

$$E_i^{T^{tc}} = \text{Min}\{\theta | (\theta \mathbf{x}_i, \mathbf{y}_i) \in T^{tc}\} \tag{1}$$

Where $(\mathbf{x}_i, \mathbf{y}_i)$ is the input/output vector for an observation i, and T^{tc} is the technology or production possibility set for year t and country c. For an input/output-vector $(\mathbf{x}, \mathbf{y})$ to be part of the production possibility set, we need to be able to produce $\mathbf{y}$ using $\mathbf{x}$. As shown in Färe and Lovell (1978), this is equivalent to the inverse of the Shephard (1970) input distance function.

If there are variable returns to scale, Farrell's measure of technical efficiency depends on the size of the observation, so that we can account for (dis)economies of scale. The measure of *technical productivity* can, following Førsund and Hjalmarsson (1987), be defined by rescaling inputs and outputs:[2]

$$E_i^{\lambda T^{tc}} = Min_{\theta,\lambda}\{\theta | (\theta \mathbf{x}_i, \mathbf{y}_i) \in \lambda T^{tc}\}, \tag{2}$$

where the convex cone of the technology λT^{tc}, contains all input–output combinations that are a proportionate rescaling of a feasible point in the technology set T^{tc}. While this is formally identical to a "CRS technical efficiency" measure, our definition here is instead that the reference surface is a homogenous envelopment of the underlying technology. This is the same assumption normally used in Malmquist indices of productivity change, see e.g. Grifell-Tatjé and Lovell (1995).

Furthermore, it is not necessary to assume that the technologies of different countries and time periods are identical in order to compare productivity, as long as one has a common reference set. It is common to use a specific (base) time period as a reference, as in Berg et al. (1992):

$$M_{ij}^{tc} = \frac{E_i^{\lambda T^{tc}}}{E_j^{\lambda T^{tc}}}, \tag{3}$$

which compares the productivity of two observations i and j using a fixed time period t as the reference, even if the observations i and j are from different time periods. A widespread alternative method is to construct geometric averages of indices based on consecutive time periods, as

[2] Førsund and Hjalmarsson (1987) used the symbols e_1 for input technical efficiency, as did Farrell (1957), and e_3 for technical productivity which they call "overall scale efficiency".

in Färe et al. (1994), which avoids the arbitrary choice of reference period t, but instead introduces a circularity problem. The approach followed here is instead to use information from all time periods for the country specific productivity reference:

$$T^c = \underset{t}{\text{Env}}(T^{tc}) \tag{4}$$

where Env() is the convex envelopment of the time specific technologies. Furthermore, to compare the productivity across countries we will need the envelopment of all time and country specific technologies:

$$\overline{T} = \underset{c}{\text{Env}}(T^c) \tag{5}$$

The reference sets (4) and (5) are not themselves technologies, only envelopment of technologies, as are the convex cones (rescaled sets) $\lambda T^c, \lambda \overline{T}$. Analogous to (2), it is then possible to define the productivity levels relative to the country specific references and the pooled references as $E_i^{\lambda T^C}$ and respectively.

The country c specific Malmquist index of productivity change over time can then be defined as.

$$M_{ij}^c = \frac{E_i^{\lambda T^c}}{E_j^{\lambda T^c}}, \tag{6}$$

which normally is reported for two observation i and j of the same unit at two points in time. In this analysis we are primarily concerned with comparing observations from different units in different countries, and there is no natural pairing of i and j. Edvardsen and Førsund (2003) develop and report geometric means of Malmquist indices between a unit in one country and all units in another country. We will instead take a simpler approach and report the productivity and efficiency *levels* of each unit and their country means.

2.2 Decomposition

As discussed e.g. in Fried et al. (2008), the Malmquist index can be decomposed in various ways, where the original decomposition is into frontier shift and efficiency change. When working in productivity and efficiency levels, the starting point is instead the decomposition of technical productivity into technical efficiency and scale efficiency:

$$E_i^{\lambda T^{tc}} = E_i^{T^{tc}} \frac{E_i^{\lambda T^{tc}}}{E_i^{T^{tc}}} = (TP_i = TE_i * SE_i), \tag{7}$$

where the parenthesis denotes the conventional way of writing the technical productivity (TP) as the product of technical efficiency ($TE_i = E_i^{T^{tc}}$) and scale efficiency ($SE_i = \frac{E_i^{\lambda T^{tc}}}{E_i^{T^{tc}}}$). By including the possibility of comparing productivity across both time and countries, this decomposition naturally expands into:

$$\begin{aligned} E_i^{\lambda \overline{T}} &= E_i^{T^{tc}} \frac{E_i^{\lambda T^{tc}}}{E_i^{T^{tc}}} \frac{E_i^{\lambda T^c}}{E_i^{\lambda T^{tc}}} \frac{E_i^{\lambda \overline{T}}}{E_i^{\lambda T^c}} \\ &= (TTP_i = TE_i * SE_i * PP_i * CP_i), \end{aligned} \tag{8}$$

where we have decomposed the now *total technical productivity (TTP)* into technical efficiency ($TE_i = E_i^{T^{tc}}$), scale efficiency ($SE_i = \frac{E_i^{\lambda T^{tc}}}{E_i^{T^{tc}}}$), period productivity ($PP_i = \frac{E_i^{\lambda T^c}}{E_i^{\lambda T^{tc}}}$) and country productivity ($CP_i = \frac{E_i^{\lambda \overline{T}}}{E_i^{\lambda T^c}}$). Each of these is specific to the observation i.

Note that dividing this decomposition for two observations of one unit at different points in time, and ignoring the country productivity, one gets the common Malmquist decomposition of technical efficiency change, scale efficiency change and frontier change. As with the Malmquist index, the decomposition is not easily extended to comparisons between countries, as there is no natural pairing of observations. Asmild and Tam (2007) develop a global index of frontier shifts which they note would be useful for international comparisons, but does not extend this to a full decomposition.

These concepts are illustrated in Fig. 1, where we ignore the time dimension and concentrate on country differences. For an observation A in country 1 with a production possibility set bounded by the production function Frontier 1, we can define the technical efficiency by (1) above as the ratio BC/BA of necessary inputs to actual inputs for a given output. The productivity of A is the slope of the diagonal OA, but we can normalise this in (2) by comparing it to the maximal productivity given by the slope of the diagonal OD. The technical productivity of A is then the ratio BD/BA. Using the definition implicit in (7), scale efficiency is

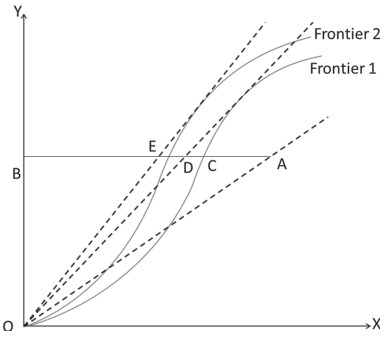


Fig. 1 The components of hospital total technical productivity in input–output space. For observation A in country 1, Total technical productivity (TTP) = BE/BA, Technical efficiency (TE) = BC/BA, Technical productivity (TP) = BD/BA, Scale efficiency (SE) = BD/BC and Country productivity (CP) = BE/BD

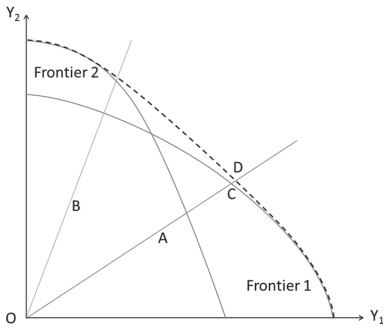


Fig. 2 The components of hospital total technical productivity in output–output space. For observation A in country 1, Total technical productivity (TTP) = OA/OD, Technical efficiency (TE) = OA/OC, and CP Country productivity (CP) = OC/OD

BD/BC. Assume that country 2 has a production possibility set bounded by Frontier 2, and that the maximal productivity of country 2 given by the slope OE is also the maximal for all countries, i.e. bounding the convex cone of all possibility sets $\lambda\bar{T}$. This slope OE will serve as the reference for the total technical productivity in (8), which for observation A is given by BE/BA. The country productivity for observation A is then the ratio BE/BD.

With only one input and one output as in Fig. 1, one country will define the reference and all observations in each country will have the same country productivity. With two outputs as in Fig. 2, the convex cone of each country's frontier λT^C can be drawn as the curved lines for a given level of the single input. The convex cone of all the country frontiers $\lambda\bar{T}$ is represented by the dashed line which serves as the reference for total technical productivity defined in (8). If the country frontiers cross as in this example, the country productivities will depend on the output mix of the observation.

2.3 Cost efficiency and productivity

Finally note that since we have only one input in our data, cost minimization for a given input price is formally equivalent to input minimization. Thus cost efficiency, which is defined as the ratio of necessary costs to input costs, is also equivalent to technical efficiency. The decomposition of productivity and the Malmquist index is most often shown in terms of technical efficiency and technical productivity but could easily have been developed in terms of cost efficiency and cost productivity. Note that in the general multi-input case the numbers will differ in technical and cost productivity decompositions, but in our one-input case, the actual numbers will be identical.[3] Thus, we may view the terms technical efficiency and cost efficiency as equivalent in discussing the results in this analysis.

2.4 Estimation method

The DEA and SFA methodologies build upon the same basic production theory basis. In both cases one estimates the production frontier (the boundry of the production possibility set or technology) or the dual formulation in the cost frontier, but the methods are quite different in their approach to estimating the frontiers and in the measures that are easily calculated and therefore commonly reported in the literature (Coelli et al. 2005; Fried et al. 2008). While the major strengths of DEA has been the lack of strong assumptions beyond those basic in theory (free disposal and convexity) and the fact that the frontier fits closely around the data, SFA has had a superior ability to handle the prescense of measurement error and to perform statistical inference. The latter shortcoming of DEA has been allieviated somewhat with the bootstrapping techniques introduced by Simar and Wilson (1998, 2000).

In our data there are good reasons to choose either method. While the prescense of measurement error is probably limited for those activities that are actually measured, there is a strong case for omitted variable (i.e. quality) bias that may be more severe in DEA. The DEA method can easily estimate the country specific frontiers without strong assumptions, thereby making country differences dependent on the input–output mix, while the SFA formulation generally introduces a constant difference between country frontiers. The prescense of country dummies in SFA implies however, that information from other countries are used to increase the precision of the estimates and therefore the power of the statistical tests.

In the DEA analysis the frontiers have been estimated using the homogenous bootstrapping algorithm from Simar and Wilson (1998), while the second stage analysis of the statistical association of technical efficiency and the environmental variables has been conducted using ordinary least square (OLS) regressions. The SFA analysis has used the simultanous estimation of the frontier component and the (in)efficiency component proposed in Battese and Coelli (1995).[4]

2.5 Data

Data has been collected for inputs and outputs of all public sector acute somatic hospitals. The hospital structure of the four Nordic countries is broadly similar. The structure

[3] In the general case with more than one output, cost efficiency and technical efficiency would be equal only if all units are allocatively efficient.

[4] The DEA bootstrap estimations have been done in FrischNonParametric, while second stage regressions and the SFA analysis has been done in STATA 12 (StataCorp 2011).

consists of mostly publicly financed and governed somatic hospitals with only a very few commercial hospitals, almost no specialization in medicine, surgical, cancer care etc., and no specialization to cater for specific groups such as veterans/military, childrens hospitals etc. Only in Finland are there a number of Health Centres with inpatient beds that serve less severe patients, and these are excluded from our analysis, as are the few commercial hospitals. Some non-profit private hospitals that are under contract with the public sector are included, however. The data includes almost the whole population of somatic hospitals in the Nordic countries, which due to a natural geographic monopoly usually serve a catchment area covering all residents. Differences in patient mix will mainly reflect demographic differences across the geographic areas, factors that are partly included in the second stage regression.

While the hospital sectors in all four countries are based on public ownership and tax-based financing, there are administrative and incentive differences. In Norway, all hospitals are state-owned, but the provision of hospital services is delegated to five (reduced to four during 2007) regional health enterprises (RHF). Each of these own between four and thirteen health enterprices (HF) which are the administrative units of hospital production, but a number of the health enterprises are multi-location institutions and the extent of integration between the actual physical hospitals varies considerably. In Denmark and Sweden hospitals are owned by the intermediate government level regions or counties ("regioner" and "landsting"), but single-location hospitals are still mainly separate institutions. The Finnish hospital sector is owned by health districts that are federations of municipalities. Norway and some counties in Sweden use partial activity based financing (ABF) based on the DRG-system, but with most of the payment made by block grants. In Denmark ABF was used only to a limited extent during the period. The Finnish hospital districts use various case-based classification systems (including DRGs) as a method of collecting payments from municipalities, but the Finnish payment system does not create similar incentives as ABF used in other countries (Kautiainen et al. 2011). However, since hospitals can be described by the same input–output vectors the productivity of the hospitals in our sample should be comparable even though they may not face the same production possibility sets.

Inputs are measured as operating costs, which for reasons of data availability are exclusive of capital costs. It was not possible to get ethical permission for the use of data for 2007 in Sweden. The Swedish data is further limited by the lack of cost information at the hospital level, nescessitating the use of the administrative county ("landsting") level as the unit of observation, each encompassing from one to five physical hospitals. The difference in level of observational unit between the countries (counties, health enterprises or hospitals) is one of the reasons why we estimate different technologies or production possibility sets in each country.

Since we do not have data on teaching and research output, the associated costs are also excluded. Costs are initially measured in nominal prices in each country's national currency, but to estimate productivity and efficiency one needs a comparable measure of "real costs" that is corrected for differences in input prices.

To harmonize the cost level between the four countries over time we have constructed wage indices for physicians, nurses and four other groups of hospital staff, as well as one for "other resources". This removes a major source of nominal cost and productivity differences between the countries, a difference that can not be influenced by the hospitals themselves, nor by the hospital sector as a whole. The wage indices are based on official wage date and include all personnel costs, i.e. pension costs and indirect labour taxes (Kittelsen et al. 2009). The index for "other resources" is the purchaser parity corrected GDP price index from OECD. The indices are weighted together with Norwegian cost shares in 2007. Thus we construct a Paasche-index using Norway in 2007 as reference point. Note that this represents an approximation, the index will only hold exactly if the relative use of inputs is constant over time and country.

Outputs are measured by using the Nordic version of the diagnosis related groups (DRGs). Each hospital discharge is assigned to one of about 500 DRGs on the basis of diagnosis and procedure codes. When activity is measured by DRG-points, discharges are weighed by a factor that is an estimate of the average cost of patients in that DRG. Thus the weighting is implicitly by patient severity or complexity as reflected in average costs. We define three broad output categories; inpatient care, day care and outpatient visits. Within each category patients are weighted with the Norwegian cost weights from 2007, where the weights are calculated from accounting data from a sample of major Norwegian hospitals.[5] Outpatient visits were not weighted. Considerable work has gone into reducing problems associated with differences in coding practice, including moving patients between DRGs, eliminating double counting etc. The problem of DRG-creep, where hospitals that face strong incentives to upcode from simple to more severe DRGs based on the number of co-morbidities has been reduced by aggregating these groups. In the DEA analysis this had the effect of reducing the mean productivity level

[5] From a common initial starting point, the Danish DRG system has diverged significantly from the other Nordic systems after 2002. Danish DRG-weights were used for the specific Danish DRG groups, while the level was normalized using those DRG-groups that were common in the two systems.

of Norwegian hospitals by 2 % points while the other countries were not affected, presumably because activity based financing is a more entrenched feature in Norway.

In addition to the single input and the three outputs, we have collected data for some characteristics that vary between hospitals within each country or over time, and that may be associated with efficiency. These include dummies for university hospital status which may capture any scope effects of teaching and research. This must be effects beyond the costs attributed to these activities which are already deducted from the cost variable, but the sign of the effect on productivity would depend on whether there are economies or diseconomies of scope between patient treatment and teaching and research. University hospitals may also have a more severe mix of patients within each DRG-group, which may bias estimated productivity downwards. The main case-mix effect should presumably already be captured by the DRG weighting scheeme. University hospitals are located in major cities. We also include a dummy for capital city hospitals, which may have a less favourable patient mix due to the socio-economic composition of the catchment area, so that one would expect the capital city hospitals to have lower productivity. However, university and capital city hospitals could also have lower costs due to shorter travelling times and a greater potential for daypatient or outpatient treatment, so the net effect is not obvious. Allthough all hospitals are located in towns, the university and capital city dummies should capture the main differences that may be due to urban or rural catchment areas.

The case-mix index (CMI) is calculated as the average DRG-weight per patient, and may again capture patient severity if the average severity within each DRG-group is correlated with the average severity as measured by the DRG-system itself, in which case one should expect a high CMI to be correlated with low productivity. The length of stay (LOS) deviation variable is calculated as the DRG-weighted average LOS in each DRG for each hospital divided by the average LOS in each DRG across the whole sample (i.e. expected LOS). Again this could capture differences in severity within each DRG group, but may also indicate excessive, and therefore inefficient, LOS. Finally, the outpatient share is an indicator of diffences in treatment practices across hospitals, where a high outpatient share may indicate lower costs per discharge. These variables are collectively termed "environmental variables", although they are not always strictly exogenous to the hospital.

In earlier studies, the extent of activity based financing (ABF) has been an important explanatory variable, but in the period covered by our dataset there has been too little variation in ABF within each country. If a variable is or highly correlated with the country then it is not possible to statistically separate the effect from other country specific fixed effects. This also holds for structural variables such as ownership structure, financing system etc. Travelling time to hospital can be an important cost driver but is not included here due to lack of data.[6] Finally, no indicators of the quality of treatment have been available for this analysis.

Table 1 shows the distribution of hospitals between countries and summary statistics for the varibles in the analyses. When interpreting the size of the Swedish observations, remember that these are not physical hospitals but the larger administrative "Landsting" units. To a lesser extent, the Norwegian observations of health enterprises can also encompass several physical hospitals.

3 Results

3.1 DEA results

In the DEA analysis, the total technical productivity level is calculated with reference to a homogenous frontier estimated from the pooled set of observations for all countries and periods. Figure 3 show that the considerable productivity superiority of the Finnish hospitals found in previous studies is also present and highly significant in this dataset. The other Nordic countries are in some periods significantly different from each other, but in general have a similar productivity level.

Figure 3 also shows a slight time trend towards declining productivity. However, the DEA bootstrap tests did not reject a hypothesis of constant technology across time periods. This implies that we can ignore the time dimension and report the simpler three-way decomposition

$$E_i^{\lambda\bar{T}} = E_i^{T^c}\frac{E_i^{\lambda T^c}}{E_i^{T^c}}\frac{E_i^{\lambda\bar{T}}}{E_i^{\lambda T^c}} = (TTP_i = TE_i * SE_i * CP_i), \qquad (9)$$

The productivity estimates for the individual observations are shown in Fig. 4. The hypothetical full productivity frontier is represented by productivity equal to 1.0, but since these numbers are bootstrapped estimates no observation is on the frontier. Clearly, the Finnish productivity level is consistently higher, with *all* Finnish observations doing better than most observations in Denmark and Norway and almost all in Sweden. Confidence intervals are quite narrow so this is a robust result. In all countries one can see that smaller units tend to be more productive, while comparisons between countries are

[6] While we do not have data for travelling time in Denmark, we have calculated the average travelling time for the catchment area of emergency hospitals in the other countries. A separate analysis reported in Kalseth et al. (2011) indicate that travelling time can explain some of the cost differences between the Norwegian regions, but not a significant amount of the differences between countries.

Table 1 Descriptive statistics. Observation means and SD

Variable	Finland	Sweden	Denmark	Norway	Total
Observation type	Hospital	Landsting/County	Hospital	Health enterprise	
Number of observations	96	40	105	75	316
Period	2005–2007	2005–2006	2005–2007	2005–2007	
Variables in production frontier function (deterministic part)					
Real Costs in billion NOK[#]	1112	4812	1516	1864	1893
SD	1563	5178	1167	1248	2488
Outpatient visits	150,128	368,134	178,620	129,609	182,321
SD	170,646	445,542	125,012	70,008	212,219
DRG points inpatients	22,516	65,262	22,517	31,447	30,047
SD	27,834	68,200	17,647	18,414	34,440
DRG points daypatients	3119	18,000	2651	4044	5067
SD	4092	18,207	2028	2532	8576
Variables in SFA efficiency part or DEA second stage (environmental variables)					
University hospital dummy	0.156	0.250	0.381	0.200	0.253
SD	0.365	0.439	0.488	0.403	0.436
Capital city dummy	0.031	0.050	0.257	0.160	0.139
SD	0.175	0.221	0.439	0.369	0.347
Case mix index DRG patients	0.848	0.655	0.915	0.918	0.862
SD	0.089	0.096	0.166	0.083	0.146
Length of stay deviation	0.968	1.118	1.017	0.859	0.977
SD	0.092	0.111	0.193	0.082	0.156
Outpatient share	0.841	0.731	0.865	0.773	0.819
SD	0.028	0.049	0.044	0.026	0.061

[#] 2007 price level

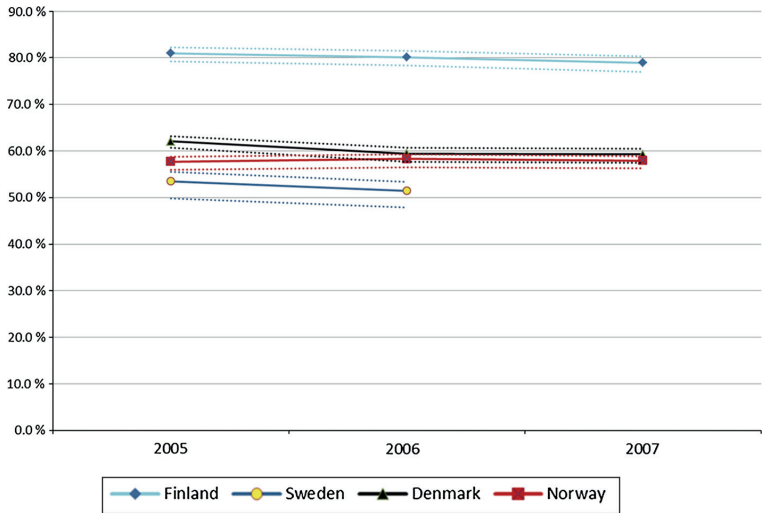


Fig. 3 DEA bootstrapped productivity estimates by country and year with common reference frontier. Mean of observations and 95 % CI

confounded by the fact that the Swedish units are not hospitals but observations on the administrative "Landsting" level.

Table 2 reports the mean country productivity results and its decomposition. The first line reports the of productivity of each country's hospital sector relative to the

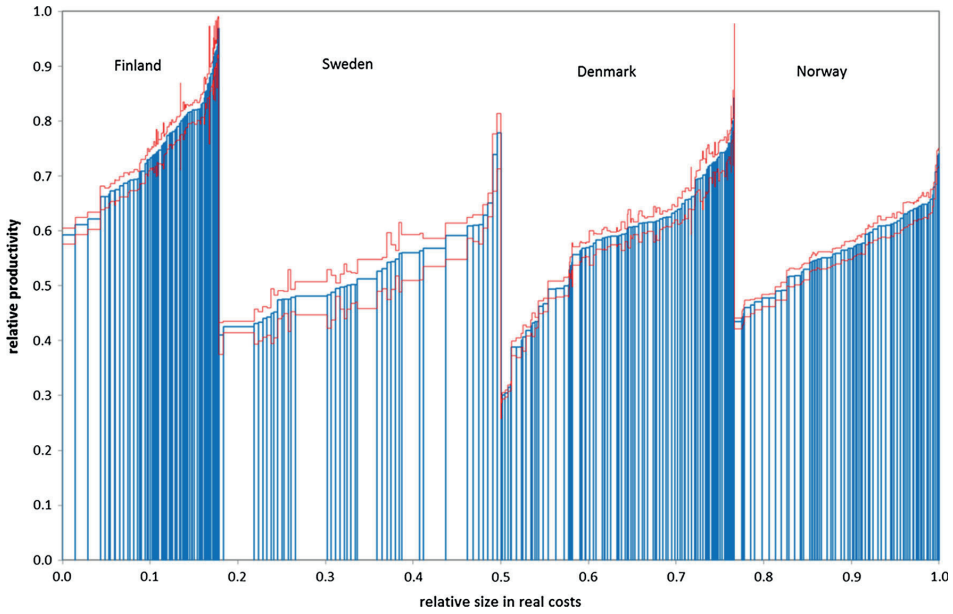


Fig. 4 Hecksher–Salter diagram of DEA bootstrapped total technical productivity estimates with pooled common reference frontier. Height of *each bar* is productivity estimate for each observation with 95 % CI, and width is proportional to the observation size measured by real costs

Table 2 Mean bootstrapped productivity in each country as measured against the pooled reference frontier in DEA

	Finland	Sweden	Denmark	Norway
Productivity with pooled reference frontier, TPP	79.1 % (77.0–81.0)	52.6 % (49.8–54.2)	57.7 % (55.4–59.6)	56.6 % (53.0–58.6)
Decomposition of productivity:				
• Productivity of country specific frontier, CP	100.0 % (99.8–100.0)	65.1 % (62.3–68.7)	78.5 % (75.8–81.4)	68.6 % (66.1–72.7)
• Scale efficiency, SE	89.7 % (87.8–91.8)	94.3 % (91.9–96.3)	93.7 % (91.9–95.2)	94.2 % (93.1–95.1)
• Technical efficiency, TE	89.8 % (88.9–90.6)	84.1 % (81.7–86.2)	77.1 % (75.4–78.6)	89.7 % (88.6–90.6)
Scale elasticity	0.935 (0.917–0.956)	1.137 (1.000–1.255)	0.940 (0.911–0.982)	0.941 (0.884–0.982)

Decomposition of total technical productivity into productivity of country specific frontier, scale efficiency and technical efficiency respectively. The mean scale elasticity is also shown

Geometric mean with 95 % CI for observations in each country

envelopment of the bootstrapped estimates of the country-specific production possibility sets, i.e. an estimate largely based on pooling the best hospitals. While Finland has an average productivity of around 80 % measured relative to the pooled frontier, the decomposition reveals that this is wholly due to lack of scale efficiency and technical efficiency, which are at around 90 % each. The country productivity mean is almost precisely 100 %, which means that it is the Finnish hospitals that define the pooled reference frontier alone.

For Sweden and Norway the picture is quite different; here the country productivity is the major component in the lack of total productivity. In fact, the cost efficiency and scale efficiency components are quite similar for Finland, Norway and Sweden. This implies that the hospitals in each country has a similar dispersion from the best to the worst performers both in terms of technical and scale efficiencies, but that the best performing hospitals in Norway and Sweden are significantly less productive than the best performers in Finland.

Denmark is in between, with significantly higher country productivity than Sweden and Norway, but still lagging far behind Finland. On the other hand, Denmark has clearly the lowest technical efficiency level of the Nordic countries, which means that the dispersion behind the frontier is largest in Denmark.

Table 3 Simplified test tree in the SFA analysis

	Log-likelihood ratio	Critical value (*df*)	Result
Should country enter the frontier function?	287.952	7.05 (3)	Yes
Is translog better than Cobb–Douglas?	42.892	11.91 (6)	Yes
Should year enter frontier function?	2.798	5.14 (2)	No
Should environmental variables enter efficiency term?	22.867	10.37 (5)	Yes
Should country enter efficiency term?	57.751	7.05 (3)	Yes
Should year enter efficiency term?	1.821	5.14 (2)	No

The Log-likelihood ratio indicator is distributed as χ^2 with degrees of freedom equal to the number of additional variables

Table 4 Marginal normalized effects on productivity in SFA and DEA, 95 % CI

Parameter	SFA		DEA	
	Frontier (deterministic component)	Efficiency component	Frontier distance	Technical efficiency in second stage regression
Finland	0.300***	0.049	0.322***	−0.029
	(0.233–0.361)	(−0.085–0.183)	(0.295–0.370)	(−0.083–0.025)
Sweden	0.071	−0.024	−0.021	−0.004
	(−0.020–0.154)	(−0.085–0.037)	(−0.068–0.061)	(−0.072–0.064)
Denmark	0.208***	−0.118***	0.050***	−0.160***
	(0.132–0.277)	(−0.174–(−0.062))	(0.010–0.094)	(−0.246–(−0.075))
Outpatient share		0.658***		0.666**
		(0.259–1.057)		(0.014–1.319)
Length of stay deviation		−0.063		−0.138**
		(−0.142–0.015)		(−0.263–(−0.013))
Case mix index		−0.048		−0.064
		(−0.160–0.064)		(−0.204–0.075)
Capital city dummy		0.030		0.040
		(−0.015–0.075)		(−0.021–1.101)
University hospital dummy		–0.010		0.012
		(−0.049–0.029)		(−0.040–0.064)
Constant		−0.216		0.533**
		(−0.504–0.072)		(0.104–1.002)

*, **, *** Significant coefficients at 10, 5 and 1 % level respectively. Reference units are hospitals in Norway in 2007 that are not in the capital and not university hospitals. The reference unit in SFA has a technical efficiency estimate of 0.9176. In the DEA model the distance between the frontiers is measured at the average product mix of Norwegian hospitals

Table 2 also reports the scale elasticities in the last line. Scale properties can be different across geographical units, as also found in a study on hospitals in two Canadian provinces by Asmild et al. (2013). Since the DEA numbers are based on separate frontier estimates for each country, the fact that the units are of a different nature represents no theoretical problem but must be reflected in the interpretation of the results. For Finland, Denmark and Norway, where the units are hospitals or low-level health enterprises, the scale elasticities below 1 indicate decreasing returns to scale on average, a result that is often found in estimates of hospital scale properties. Thus, optimal size is smaller than the median size. For Sweden, however, the scale elasticity is larger than one, although only just significantly. Thus, even though the units of observation are clearly larger in Sweden, the optimal size is even larger. The natural interpretation of this paradox is that while the optimal size of a hospital is quite small, the optimal size of an administrative region (or purchaser), such as the Swedish Landsting, is quite large. Of course, other national differences that are not captured by our variables may also explain this result.

3.2 SFA results

The testing tree for the SFA model is shown in Table 3. The formulation by Battese and Coelli (1995) implies that factors that determine the position of the frontier function in the deterministic part of the equation are estimated

simultaneously as the variables in the "explanation" of the inefficiency term. Right hand side variables can potentially enter both components.

Clearly, the strongest result is that country dummies should enter the frontier term. This implies that there are highly significant fixed country effects that are not explained by any of our other variables, and that by the assumptions of the model specification the country dummy should primarily shift the frontier term. The functional form of the inefficiency term is not easily tested but the exponential distribution is the one that fits the data most closely. The functional form of the frontier function itself is, however, testable, and the simple Cobb-Douglas form is rejected in favour of the flexible Translog form. The time period dummies are also rejected in both terms, which mean that the period can be ignored as in the DEA case.

The normalized marginal effects are shown in Table 4 together with the corresponding DEA results. The full estimation results for the preferred model are included in Appendix 2. The normalization in Table 4 is done so that a positive coefficient shows the percentage point increase in the productivity level (or decrease in costs) stemming from a one per cent increase in the explanatory variable. The frontier and efficiency terms are shown in separate columns. For the DEA results, the marginal effects are dependent on the input–output mix, and the numbers shown are for the average Norwegian observation.

The results are generally very robust across methods. The Finnish hospitals are strongly more productive than the other countries. The Swedish and Norwegian frontiers are not significantly different from each other, while the Danish frontier is in between the Finnish and the Swedish/Norwegian. In the efficiency term, the only significant country effect is that the Danish hospitals are less efficient. Of the environmental variables, the outpatient share has a significant positive effect on productivity while the LOS deviation has a weaker negative effect. The case-mix index and the dummies for university and capital city hospitals have no effect on costs. There seems to be no sign that the central hospitals have a more costly case mix than what is accounted for by the DRG system.

4 Conclusion

International comparisons can reveal more about the cost and productivity structure of a sector such as the somatic hospitals than a country specific study alone. In addition to an increase in the number of observations and therefore in the degrees of freedom, one gets more variation in explanatory variables and stronger possibilities for exploring causal mechanisms. This study has found evidence of a positive association between efficiency and outpatient share, a negative association with LOS, and no association with the case-mix index or university and capital city dummies. We have further found evidence of decreasing returns to scale at the hospital level, with a possibility of increasing returns to scale at the administrative or purchaser level. There is also evidence of cost/technical inefficiency, particularly in Denmark.

As so often, the strongest results are not what we can explain, but what we cannot explain. There is strong evidence, independent of method, that there are large country specific differences that are not correlated with any of our other variables. Finland is consistently more productive than the other Nordic countries. There are systematic differences between countries that do not vary between hospitals within each country. Without observations from more countries, or more variables that vary over time or across hospitals within each country, such mechanisms cannot be revealed by statistical methods.

On the other hand, qualitative information can give some speculations and plausible explanations. Interestingly, the stronger incentives that are supposed to be provided by ABF in Norway and some counties of Sweden does not seem to increase productivity. These data are from before the financial crisis, but Finland was still suffering the aftereffects of a local recession after the collapse of the Soviet Union, with increased budget restraint in the public sector. Based on interviews of 8 hospitals in Nordic countries (Kalseth et al. 2011) some of the possible reasons for the Finnish good results can be the good coordination between somatic hospitals and primary care, including inpatient departments of health centres. This coordination is primarily due to the common ownership by the municipalities of both hospitals and primary care institutions.[7] Finland also had a smaller number of personnel as well as better organization of work and team work between different personnel groups inside hospitals (Kalseth et al. 2011). However, these findings are still preliminary. An important research and policy question is whether the higher productivity in Finland is related to differences in quality.

Our claim is that the country productivity differences are consistent with possible differences in system characteristics that vary systematically between countries. Such characteristics may include the financing structure, ownership structure, regulation framework, quality differences, standards, education, professional interest groups, work culture, etc. Some of these characteristics, such as quality,

[7] As mentioned, Finland has low-speciality health centres that are excluded from study. If these treat the least severe patients then the Finnish hospitals would have a more severe case-mix. Most of this should be captured in the DRG-system, but if hospital patients are more severe within each DRG the potential bias is that the Finnish hospitals are actually even more productive than estimated here.

may also vary between hospitals in each country and should be the subject of further research.

Differences in estimated country productivity are also consistent with data definition differences, but the analysis in Kalseth et al. (2011) does not support this. In summary, these country effects are essentially not caused by factors that can be changed by the individual hospitals to become more efficient, but rather factors that must be tackled by relevant organizations and authorities at the national level.

Acknowledgments We acknowledge the contribution of other participants in the Nordic Hospital Comparison Study Group (http://www.thl.fi/en_US/web/en/research/projects/nhcsg) in the collection of data and discussion of study design and results. During this study the NHCSG consisted of Mikko Peltola and Jan Christensen in addition to the authors listed. The data has been processed by Anthun, with input from Kalseth and Hope, while Kittelsen and Winsnes have performed the DEA and SFA analysis respectively and drafted the manuscript. All authors have critically reviewed the manuscript and approved the final version. We thank the Norwegian board of health and the Health Economics Research Programme at the University of Oslo (HERO—www.hero.uio.no), the Research Council of Norway under grant 214338/H10, as well as the respective employers, for financial contributions. We finally thank the participants of the Conference in Memory of Professor Lennart Hjalmarsson in December 2012 in Gothenburg for helpful comments and suggestions.

Appendix 1: Estimated models

Let x_i denote the level of the single input for each hospital i and $\mathbf{y}_i$ is the vector of y_{ni}, the level of output n for hospital i. The environmental variables that enter the frontier function are denoted z^a_{ki}, and the environmental variables that enter the (in)efficiency function z^b_{li}. After the tests described in Table 3, z^a_{ki} only consists of country dummies (with Norway as the dropped reference), while $z_{li}{}^b$consists of all variables in the first column of Table 4.

The set of observed hospitals i in each country c and period t is denoted H_{ct}, the intertemporal set in each country is $H_c = \bigcup_t H_{ct}$, and the full set of all observations across all countries is $\bar{H} = \bigcup_c H_c$.

DEA estimates

For an input–output vector $(x_0, \mathbf{y}_0)$, the basic estimates of country-and period-specific technical input efficiencies used in Table 1 are bootstrapped estimates using the standard DEA variable returns to scale (VRS) formulation:

$$E_0^{T^{tc}} = \text{Min}\left\{ \theta \middle| \theta x_0 \geq \sum_{i\in H^{tc}} \lambda_i x_i,\ \mathbf{y}_0 \leq \sum_{i\in H^{tc}} \lambda_i \mathbf{y}_i,\ \sum_{i\in H^{tc}} \lambda_i = 1 \right\} \tag{10}$$

Where the restrictions represent the DEA estimate of the production possibility set T^{tc}. The period- and country-specific *technical productivity* is then the measured relative to the homogenous envelopment $\lambda T^{tc}$of the set:

$$E_0^{\lambda T^{tc}} = \text{Min}\left\{ \theta \middle| \theta x_0 \geq \sum_{i\in H^{tc}} \lambda_i x_i,\ \mathbf{y}_0 \leq \sum_{i\in H^{tc}} \lambda_i \mathbf{y}_i \right\} \tag{11}$$

The estimate of productivity of an observation relative to the intertemporal country-specific frontier is calculated relative to a reference set pooling all observations in each country across periods:

$$E_0^{\lambda T^{c}} = \text{Min}\left\{ \theta \middle| \theta x_0 \geq \sum_{i\in H^{c}} \lambda_i x_i,\ \mathbf{y}_0 \leq \sum_{i\in H^{c}} \lambda_i \mathbf{y}_i \right\} \tag{12}$$

The estimate of productivity of an observation relative to the intertemporal and cross-country frontier, i.e. the *total technical productivity* is calculated relative to a reference set pooling all observations across all countries and periods:

$$E_0^{\lambda \bar{T}} = \text{Min}\left\{ \theta \middle| \theta x_0 \geq \sum_{i\in \bar{H}} \lambda_i x_i,\ \mathbf{y}_0 \leq \sum_{i\in \bar{H}} \lambda_i \mathbf{y}_i \right\} \tag{13}$$

After elimination of the assumption of period-specific frontiers from the decomposition (9), technical efficiency is calculated with the intertemporal pooled country reference sets:

$$E_0^{T^{c}} = \text{Min}\left\{ \theta \middle| \theta x_0 \geq \sum_{i\in H^{c}} \lambda_i x_i,\ \mathbf{y}_0 \leq \sum_{i\in H^{c}} \lambda_i \mathbf{y}_i,\ \sum_{i\in H^{c}} \lambda_i = 1 \right\} \tag{14}$$

All estimates are bootstrapped using the homogenous procedure in Simar and Wilson (1998), with 2000 bootstrap iterations and kernel estimates of the inefficiency distributions based on the technical efficiency scores (10) and (14) respectively.

The second stage regression in the DEA analysis is an OLS regression with bootstrapped technical efficiency estimates $\hat{E}_i^{T^c}$ for each hospital i as the dependent variable and environmental variables z^b_{li}as independent variables of the form:

$$\hat{E}_i^{T^c} = \gamma_0 + \sum_{l=1}^{L} \gamma_l z^b_{li} + \varepsilon_i \tag{15}$$

SFA estimates

The stochastic frontier analysis is based on maximum likelihood estimation of the Battese and Coelli (1995) type

of model with the variance of the inefficiency term is a function of the environmental variables. The model is estimated using the "cost function" procedure in Stata 13 as our only input is total operating costs. After the tests in Table 3, the reported results are from a model with a translog functional form for the deterministic part and exponential distribution for the inefficiency term.

$$\ln x_i = \beta_0 + \sum_{n=1}^{N} \beta_n \ln y_{ni} + 0.5 \sum_{n=1}^{N} \sum_{m=1}^{N} \beta_{nm} \ln y_{ni} \ln y_{mi} + \sum_{k=1}^{K} \eta_k z_{ki}^1 + v_i - u_i v_i \sim Normal(0, \sigma_v^2)$$
$$u_i \sim Exponential(\sigma_{ui}),\ \ln(\sigma_{ui}^2) = \delta_0 + \sum_{l=1}^{L} \delta_l z_{li}^b \qquad (16)$$

Appendix 2: Raw coefficients in SFA analysis

See Table 5.

Table 5 SFA-analysis with an exponentially distributed efficiency component

	Coefficient	Z-value
Cost frontier (deterministic part)		
Constant	−10.349	−9.760***
Ln outpatients	0.410	0.930
Ln DRG inpatients	0.093	0.260
Ln DRG daypatients	0.500	2.090**
(Ln Outpatients)* (Ln DRG inpatients)	−0.246	−2.030**
(Ln Outpatients)* (Ln DRG daypatients)	0.030	0.510
(Ln DRG inpatients)* (Ln DRG daypatients)	−0.240	−2.690***
(1/2) (Ln Outpatients)^2	0.193	1.940*
(1/2) (Ln DRG inpatients)^2	0.530	2.760***
(1/2) (Ln DRG daypatients)^2	0.200	3.650***
Finland	−0.356	−7.660***
Sweden	−0.074	−1.540
Denmark	−0.233	−4.970***
Inefficiency part		
Constant	5.457	1.470
Finland	−1.236	−0.720
Sweden	0.610	0.770
Denmark	2.984	4.120***
Outpatient share	−16.634	−3.230***
Length of stay deviation	1.602	1.580
Case mix index	1.214	0.840
Capital city dummy	−0.755	−1.290
University hospital dummy	0.244	0.480
Log likelihood	218.275	
Scale elasticity	0.928	
Gradient vector	2.31e-7	
Number of observations	316	
Number of regular observations	249	

Dependent variable is total real costs in billion 2007 NOK. Reference units are hospitals in Norway in the year 2007, which is neither in the capital nor are university hospitals

*, **, *** Significant coefficients at 10, 5 and 1 % level respectively. In the inefficiency part, positive coefficients indicate reduced efficiency. Scale elasticity is calculated as in Coelli et al. (2005). Regularity conditions for the cost frontier part are as calculated in Salvanes and Tjotta (1998)

References

Asmild M, Tam F (2007) Estimating global frontier shifts and global Malmquist indices. J Prod Anal 27:137–148. doi:10.1007/s11123-006-0028-0

Asmild M, Hollingsworth B, Birch S (2013) The scale of hospital production in different settings: one size does not fit all. J Prod Anal. doi:10.1007/s11123-012-0332-9

Battese GE, Coelli TJ (1995) A model for technical inefficiency effects in a stochastic frontier production function for panel data. Empir Econ 20:325–332

Berg SA, Førsund FR, Jansen ES (1992) Malmquist indices of productivity growth during the deregulation of Norwegian banking 1980–1989. Scand J Econ 94:211–288

Caves DW, Christensen LR, Diewert WE (1982) The economic theory of index numbers and the measurement of input, output, and productivity. Econometrica 50:1393–1414

Coelli TJ, Rao DSP, O'Donnell CJ, Battese GE (2005) An introduction to efficiency and productivity, vol 2. Springer, Verlag

Derveaux B, Ferrier GD, Leleu H, Valdmanis V (2004) Comparing French and US hospital technologies: a directional input distance function approach. Appl Econ 36:1065–1081

Edvardsen DF, Førsund FR (2003) International benchmarking of electricity distribution utilities. Res Energ Econ 25:353–371

Färe R, Lovell CAK (1978) Measuring the technical efficiency of production. J Econ Theor 19:150–162

Färe R, Grosskopf S, Lindgren B, Roos P (1994) Productivity developments in Swedish hospitals; a Malmquist output index approach. In: Charnes A, Cooper W, Lewin AY, Seiford LM (eds) Data envelopment analysis: theory, methodology and applications. Kluwer Academic Publishers, Massachusets, pp 253–272

Farrell MJ (1957) The measurement of productive efficiency. J R Stat Soc 120:253–281

Førsund FR, Hjalmarsson L (1987) Analyses of industrial structure: a putty-clay approach. Almqvist and Wiksell International, Stockholm

Fried HO, Lovell CAK, Schmidt SS (2008) The measurement of productive efficiency and productivity growth. Oxford University Press, Oxford

Grifell-Tatjé E, Lovell CAK (1995) A note on the Malmquist productivity index. Econ Lett 47:169–175

Halsteinli V, Kittelsen SA, Magnussen J (2010) Productivity growth in outpatient child and adolescent mental health services: The

impact of case-mix adjustment. Soc Sci Med 70:439–446. doi:10.1016/j.socscimed.2009.11.002

Kalseth B, Anthun KS, Hope Ø, Kittelsen SAC, Persson B (2011) Spesialisthelsetjenesten i Norden. Sykehusstruktur, styringsstruktur og lokal arbeidsorganisering som mulig forklaring på kostnadsforskjeller mellom landene. SINTEF Report A19615, SINTEF Health Services Research, Trondheim

Kautiainen K, Häkkinen U, Lauharanta J (2011) Finland: DRGs in a decentralized health care system. In: Busse R, Geissler A, Quentin W, Wiley M (eds) Diagnosis-related groups in Europe: Moving towards transparency, efficiency and quality in hospitals. European Observatory on Health Systems and Policies Series. McGraw-Hill, Maidenhead, pp 321–338

Kittelsen SAC, Anthun KS, Kalseth B, Halsteinli V, Magnussen J (2009) En komparativ analyse av spesialisthelsetjenesten i Finland, Sverige, Danmark og Norge: Aktivitet, ressursbruk og produktivitet 2005–2007. SINTEF Report A12200, SINTEF Health Services Research, Trondheim

Kittelsen SAC, Magnussen J, Anthun KS, Häkkinen U, Linna M, Medin E, Olsen K, Rehnberg C (2008) Hospital productivity and the Norwegian ownership reform—a Nordic comparative study. STAKES discussion paper 2008:8, STAKES, Helsinki

Linna M, Virtanen M (2011) NordDRG: the benefits of coordination. In: Busse R, Geissler A, Quentin W, Wiley M (eds) Diagnosis-related groups in Europe: moving towards transparency, efficiency and quality in hospitals. Open University Press, Maidenhead

Linna M, Hakkinen U, Magnussen J (2006) Comparing hospital cost efficiency between Norway and Finland. Health Policy 77:268–278. doi:10.1016/j.healthpol.2005.07.019

Linna M, Häkkinen U, Peltola M, Magnussen J, Anthun KS, Kittelsen S, Roed A, Olsen K, Medin E, Rehnberg C (2010) Measuring cost efficiency in the Nordic Hospitals-a cross-sectional comparison of public hospitals in 2002. Health Care Manag Sci 13:346–357. doi:10.1007/s10729-010-9134-7

Magnussen J (1996) Efficiency measurement and the operationalization of hospital production. Health Serv Res 31:21–37

Malmquist S (1953) Index numbers and indifference surfaces. Trabajos de estadistica 4:209–224

Medin E, Häkkinen U, Linna M, Anthun KS, Kittelsen SAC, Rehnberg C (2013) International hospital productivity comparison: experiences from the Nordic countries. Health Policy 112:80–87. doi:10.1016/j.healthpol.2013.02.004

Mobley L, Magnussen J (1998) An international comparison of hospital efficiency. Does institutional environment matter? Appl Econ 30:1089–1100

Salvanes KG, Tjøtta S (1998) A note on the importance of testing for regularities for estimated flexible functional forms. J Prod Anal 9:133–143

Shephard RW (1970) Theory of cost and production functions, 2nd edn. Princeton University Press, Princeton

Simar L, Wilson PW (1998) Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. Manag Sci 44:49–61

Simar L, Wilson PW (2000) Statistical inference in nonparametric frontier models: the state of the art. J Prod Anal 13:49–78

StataCorp (2011) Stata: Release 12. Statistical Software. StataCorp LP, College Station

Steinmann L, Dittrich G, Karmann A, Zweifel P (2004) Measuring and comparing the (in)effciency of German and Swiss Hospitals. Eur J Health Econ 5:216

Varabyova Y, Schreyögg J (2013) International comparisons of the technical efficiency of the hospital sector: panel data analysis of OECD countries using parametric and non-parametric approaches. Health policy. doi:10.1016/j.healthpol.2013.03.003

Paper III

HEALTH ECONOMICS
Health Econ. **24**(Suppl. 2): 140–163 (2015)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/hec.3260

COSTS AND QUALITY AT THE HOSPITAL LEVEL IN THE NORDIC COUNTRIES

SVERRE A. C. KITTELSEN[a,*], KJARTAN S. ANTHUN[b], FANNY GOUDE[c], INGRID M. S. HUITFELDT[a], UNTO HÄKKINEN[d], MARIE KRUSE[e], EMMA MEDIN[c], CLAS REHNBERG[c], HANNA RÄTTÖ[d]
ON BEHALF OF THE EUROHOPE STUDY GROUP

[a]*Frisch Centre, Oslo, Norway*
[b]*SINTEF Health Research and NTNU, Trondheim, Norway*
[c]*Medical Management Centre, Karolinska Institutet, Stockholm, Sweden*
[d]*Centre for Health and Social Economics CHESS, National Institute for Health and Welfare, Helsinki, Finland*
[e]*COHERE, University of Southern Denmark, Odense, Denmark*

ABSTRACT

This article develops and analyzes patient register-based measures of quality for the major Nordic countries. Previous studies show that Finnish hospitals have significantly higher average productivity than hospitals in Sweden, Denmark, and Norway and also a substantial variation within each country. This paper examines whether quality differences can form part of the explanation and attempts to uncover quality–cost trade-offs.

Data on costs and discharges in each diagnosis-related group for 160 acute hospitals in 2008–2009 were collected. Patient register-based measures of quality such as readmissions, mortality (in hospital or outside), and patient safety indices were developed and case-mix adjusted. Productivity is estimated using bootstrapped data envelopment analysis.

Results indicate that case-mix adjustment is important, and there are significant differences in the case-mix adjusted performance measures as well as in productivity both at the national and hospital levels. For most quality indicators, the performance measures reveal room for improvement. There is a weak but statistical significant trade-off between productivity and inpatient readmissions within 30 days but a tendency that hospitals with high 30-day mortality also have higher costs. Hence, no clear cost–quality trade-off pattern was discovered. Patient registers can be used and developed to improve future quality and cost comparisons. Copyright © 2015 John Wiley & Sons, Ltd.

Received 5 March 2014; Revised 27 April 2015; Accepted 11 May 2015

KEY WORDS: international comparisons; quality; outcomes; performance; productivity

1. INTRODUCTION

Increasing health expenditures and a growing demand for health services have put an increasing focus on cost containment and the efficiency of delivering health services. However, the pressure to contain costs through enhanced efficiency may lead to poorer quality (Gutacker *et al*., 2013), which emphasizes the need for controlling for quality. Low quality could also be linked to high wasteful costs (McKay and Deily, 2008). Previous studies investigating the relationship between costs and quality show conflicting findings and use of heterogeneous methods and measures (Hussey *et al*., 2013). Hence, more knowledge on the association between provider costs and treatment quality is needed, and the use of cross-country comparisons gives opportunities to identify similarities and differences (Häkkinen *et al*., 2015).

An important issue in exploring the association between quality and costs is the choice of quality indicators. The indicators must reflect aspects that are of value to patients or society, which imply, in the nomenclature of

*Correspondence to: Frisch Centre, Gaustadalléen 21, NO-0349 Oslo, Norway. E-mail: sverre.kittelsen@frisch.uio.no

Donabedian (1966), that they should be at least process or structural quality indicators that are related to outcomes. To be useful at the hospital level, the indicators should be relevant for a non-negligible portion of the patients and be able to statistically distinguish between hospitals. The most interesting measures would be those that reflect medical quality by showing improved health, but from an economics point of view, measures of service quality could also be relevant if they reflect aspects of value to patients.

Several studies relate hospital costs to in-hospital or post-hospitalization mortality rates (Hussey *et al*., 2013). In-hospital mortality used as the main quality indicator however poses some challenges. On the one hand, treatment costs could be low if the patient dies quickly after the admission. On the other hand, many resources are used for patients during their last days before death. This means that costs are endogenous to health outcomes, but these problems are less severe when mortality is measured regardless of death occurring in hospital or after discharge, as in this study. Alternative indicators may be based on complications (e.g., Kruse and Christensen, 2013), which however can be quite procedure specific and hence difficult to compare across medical specialities. Readmission rates encompass aspects of both medical and service quality.

The literature suggests that there will be a U-shaped relation between costs and quality, which for higher levels of quality means that there is a trade-off between cost containment and quality improvement, while for lower levels of quality, there may be a cost-saving potential of quality improvements (Hvenegaard *et al*., 2009; Carey and Stefos, 2011; Gryna, 1999; Hvenegaard *et al*., 2011). The intuition of the U-shaped relation would be that at lower quality levels, investments for improving quality may lower the net cost of treatment. Meanwhile, hospitals at higher levels of quality may operate on the upward sloping part where further investments may improve quality. If hospital service production is efficient, there will be a trade-off between quality and quantity or equivalently between costs and quality. All other things being equal, one cannot then increase the quality of treatment without incurring some opportunity costs such as reducing the number of patients treated or alternatively using more resources.

In empirical cross-section studies that compare hospitals, the relation will often be negative (e.g., Kruse and Christensen, 2013). This could be because of inadequate case-mix adjustment because some patients are inherently more prone to (costly) complications and readmissions and therefore have higher expected costs. Also, if the number of cases is small, there could be a large random component in the likelihood of complications. If case-mix adjustment is adequate and the number of cases is sufficient to disregard random variations, there remains the possibility of inefficiency. If it is possible to improve quality without increasing costs or reducing quantity, then the treatment is inefficient. On a more positive note, it is possible that a hospital that provides good quality may also be good at containing costs.

In their study of cost inefficiency and mortality in Florida hospitals, Deily and McKay (2006) isolated costs due to inefficiency and found a strong association to mortality. Their study applied individual level data in a stochastic frontier analysis. In a later study including a later sample of US acute hospitals, the authors found no systematic pattern of association between cost inefficiency and hospital outcome (McKay and Deily, 2008). Carey and Burgess (1999) found a positive relationship between costs and outpatient follow-up within 30 days after inpatient discharge for a sample of Veterans Administration (VA) hospitals in the USA. Fleming (1991) analyzed the cost and mortality/readmission relationship for Medicare beneficiaries hospitalized at 659 US hospitals and found that higher cost had a cubic association with the readmission index and surgical mortality index. Total and medical mortalities were not significantly associated with cost. Morey *et al*. (1992) used a national sample of 350 US hospitals to analyze the relationship between data envelopment analysis (DEA) scores and actual to predicted in-hospital deaths. They found that a reduction of one death was associated with an increase in efficient cost of $28,926. Mukamel *et al*. (2001) found a positive relationship between costs and risk-adjusted 30-day mortality after discharge for Medicare beneficiaries.

In a recent Canadian study, Stukel *et al*. (2012) found a positive association between costs and quality in a longitudinal analysis at patient level. They analyzed the association of hospital spending intensity and mortality and readmission rates for four common conditions (acute myocardial infarction (AMI), chronic heart failure, hip fracture, and colon cancer) in 129 hospitals in Ontario. This finding was confirmed by a German study, also at patient level, where they examined health outcomes (mortality at 30, 60, 90, and 365 days after discharge) for

DOI: 10.1002/hec

AMI as a function of costs and other patient-level variables in 318 German hospitals (Stargardt *et al*., 2014). Birkmeyer *et al*. (2012) examined the relationships between hospital outcomes (complication rates at inpatient surgery) and risk-adjusted, 30-day episode payments for four acute and elective procedures in US hospitals. It appeared that the complication rate was positively associated with Medicare payments, indicating a negative association between costs and quality. There was no statistical significant association between costs and mortality, however.

The survey by Hussey *et al*. (2013) attributed the divergent conclusions on the cost–quality association partly to differences in the unit of analysis (hospital, department, or patient group), measurement of costs and quality, as well as the adapted methodology. Hospital studies were slightly more likely to report a positive association between costs and quality than studies using other levels (such as nursing homes or areas) of analysis.

Studies under the EuroHOPE project have made advances in the comparison of healthcare costs between countries and relate the costs to outcomes and quality (e.g., Iversen *et al*., 2015; Heijink *et al*., 2015), but these studies look at a restricted set of diagnoses at a time. A recent study of the Organisation for Economic Cooperation and Development (OECD) countries analyzed the association between costs and efficiency for hospitals as a whole (Varabyova and Schreyögg, 2013). This article aims to expand such comparisons to include the quality of care as well, measured by selected case-mix adjusted quality variables. While this study relates to the EuroHOPE project, it includes only the four major Nordic countries (Norway, Sweden, Finland, and Denmark) in the comparison because only these countries have nationwide patient registers applicable for usage of the same hospital-wide case-mix (diagnosis-related group (DRG)) system. The homogenous definition of hospital outputs used in patient registers in the Nordic countries facilitates fair comparisons across countries.

Previous studies have indicated that Finnish hospitals have significantly higher average productivity than hospitals in Sweden, Denmark, and Norway and a substantial variation within each country (Kittelsen *et al*., 2008; Linna *et al*., 2010; Medin *et al*., 2011; Kalseth *et al*., 2011). Controlling for within-country variations in activity-based financing, length of stay (LOS), outpatient shares, university hospital status, or capital region only contributes to a small portion of these differences.

This paper examines whether quality differences can form part of the explanation for productivity differences and attempts to uncover any quality–cost trade-off at the hospital level. The analysis uses both individual patient-level and hospital-level data while taking cross-country differences into account. Auxiliary aims are to evaluate the usefulness of available quality indices and the importance of case-mix adjustments in these analyses. The pooling of data from four countries has at least two advantages. Firstly, we have a much larger sample size; and secondly, we are able to identify whether our findings are due to nation-specific or structural factors.

2. DATA

To perform the analysis in this study, we use data on hospital input and both quantitative and qualitative outputs. The productivity analysis utilizes a single input of hospital costs and three DRG-weighted outputs (medical inpatients, surgical inpatients, and outpatient visits) based on patient-level discharge registry data from 2008 to 2009. Individually identifiable patient data were not available in Norway before 2008. To calculate 365-day mortality, demographic data have been collected also for 2010. The Danish data are affected by the strike among hospital nurses in non-acute functions in 2008. Although one might expect a productivity penalty from the strike, both DRG production and costs would be reduced, and the impact on productivity should be minor. This section describes the hospital costs and patient-level discharge data sets, their sources and definitions, as well as the quality indicators and the case-mix adjustment variables used in the analysis (more details are available in Medin *et al*. (2013) and Anthun *et al*. (2013)). In the study, only somatic hospitals with a 24-hour emergency department or at least two medical or surgical specialities are included.

DOI: 10.1002/hec

2.1. Cost data

The hospitals costs include all production-related costs from the hospitals. Costs were harmonized across the countries by excluding costs for ambulances, value added tax (VAT), capital costs, purchased care, and costs for teaching and research.[1]

In Sweden, the cost data were assembled mainly from the Swedish Association of Local Authorities and Regions through the cost per patient database, from hospital annual reports, and from Statistics Sweden. The hospitals not recorded in these sources were sent a cost survey. For six Swedish counties, it was not possible to create data at the hospital level; so for these counties,[2] the output was also aggregated to the county level.

In Norway, the cost data were derived from the SAMDATA database of Norwegian specialized care published annually by the Directorate of Health. The National Institute for Health and Welfare in Finland collects hospital cost data annually as part of hospital productivity statistics production, while annual productivity reports published by the Ministry of Health contained the Danish cost data.

2.1.1. Cost level deflator. The collected cost data were measured in nominal prices in each country, and the costs were deflated to create real costs in each country. There were differences in currencies and input prices between the countries, and to allow for comparison between countries, the cost level had to be harmonized.

Wage indices were calculated for nine of the most important personnel groups. The wage indices were based on official wage data for the nine separate groups and included all personnel costs such as wage taxes and pension contributions (Anthun *et al*., 2013; Kittelsen *et al*., 2009; Medin *et al*., 2013). Personnel costs are the most important component with about 60% of total hospital costs. For the other costs, we use the purchasing power parity-adjusted gross domestic product price index from OECD. To form the aggregate cost level deflator, the nine personnel group indices and the index for other costs were weighed with fixed Norwegian shares for 2008, as personnel shares were not available for the other countries.

2.2. Patient-level data

Patient-level data were collected from national administrative patient registries in all four countries. The level of data was departmental (speciality) discharges. Outpatient visits registered during inpatient stays were excluded.

Death outside of hospitals was collected by linking patient-level data with other registries. In Norway, this linkage is automatically carried out in the patient registry through a link with the National Population Registry. The Danish patient data were manually linked with the Population Register. In Sweden and Finland, the time to death was collected by manually linking with the cause of death registries.

2.2.1. Diagnosis-related group grouping and weights. Norway, Sweden, and Finland each have a national version of a common grouping system for the hospital visits, NordDRG, developed at the Nordic Casemix Centre.[3] Denmark used to be part of NordDRG but changed to a national system DkDRG in 2002. The DkDRG system applies similar rules but is not completely comparable at the DRG level (Medin *et al*., 2013). Even though three of the countries have highly comparable systems, a common grouping is to be desired in order to enhance the comparability of the output measures and quality indices and to remove some of the idiosyncrasies inherent in each health system. All four countries have patient registers that use the same diagnosis and procedural classification systems, and Datawell Oy Finland has developed a common Nordic grouper for use in this and other projects based on definitions from the Nordic Casemix Centre. This grouper allows for similar grouper logic to be applied to all four Nordic countries. All patient-level data were regrouped in this grouper.

[1]Some additional costs were also excluded, details available in Anthun *et al*. (2013).
[2]Blekinge, Västmanland, Jämtland, Dalarna, Gävleborg, and Värmland. Kronoberg, Södermanland and Gotland have additionally been excluded from the productivity analysis because of problems in the cost data.
[3]http://www.nordcase.org/

DOI: 10.1002/hec

Common DRG weights are also needed to compare the countries. A set of cost weights were calculated from pooled 2008 and 2009 cost per patient data from Helsinki and Uusimaa hospital district in Finland grouped with the common Nordic grouper. As a robustness exercise, we have also calibrated weights for each of the Nordic DRGs using the average Swedish DRG weights of the Swedish patients assigned to that Nordic DRG.

Table I. Definitions of variables used

Group	Variable name	Definition
Quality indicators		
Readmissions	Readm30_Emergency	Patient admitted acutely to inpatient care in hospital within 30 days of the discharge
	Readm30_Inpatient	Patient admitted to inpatient care in hospital within 30 days of the discharge and at least two days after discharge
Mortality	Mort30_LastAdmittance	Out of hospital mortality from any cause. Dummies
	Mort90_LastAdmittance	for 30, 90, 180 and 365 days after admission.
	Mort180_LastAdmittance	Only set for last admission within the specified period.
	Mort365_LastAdmittance	
Patient safety indicators		PSI indicators as defined by OECD
	PSI12_vt_pe	Pulmonary embolism/Deep vein thrombosis
	PSI13_Sepsis	Sepsis
	PSI15_AccidCutPunc	Accidental cut, puncture, or haemorrhage during medical care
	PSI18_ObstTrauma	Obstetric trauma
	BedSores	Bed-sores
Case-mix adjustment variables (used in models 0-5)		
Model 0: Nordic DRG	DRG	Diagnosis related group based upon common Nordic grouper
Model 1: + Patient characteristics		
	Male	1=Male, 0=Female
	Agegrp0	Age dummies for the groups:
	Agegrp1_9	0, 1-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69,
	…	70-79, 80-89, 90+
	Agegrp80_89	
	Agegrp90	
Model 2: + Treatment variables		
	TransInOwnHospital	Dummies for transfer into and out of hospital department
	TransInOtherHospital	stay within one day before or after this stay.
	TransOutOwnHospital	Not based on original coding but calculated from dates
	TransOutOtherHospital	of patient registry directly.
	Charlson	Charlson index based upon secondary diagnosis
	NumSecDiagnoses	Number of secondary diagnoses
Model 3: + Length of stay	LOS	Length of stay defined as discharge date – admission date + 1
Model 4: + Municipal variables for patient	Population	Population of patient home municipality
	Unemployment	Unemployment rate as % of labour force
	SocialAssist	Social assistance recipients as % of population
	SingleFamilies	Single parent families, as % of all families with children
	Foreign	Citizens of foreign countries as % of population
Model 5: + Hospital-municipal variable	Traveltime	Travelling time by car in hours between hospital and centre of home municipality
Hospital level variables		
	Costs	Deflated real operating costs in common currency, corrected for differences in input price level between countries and years.
	Average Costs	Costs divided by total DRG-points
	NumberOfPatients	Number of departmental/speciality discharges
	Case-mix index (CMI)	Hospital DRG-points divided by number of patients
	UniversityHospital	Dummy if hospital is a teaching or university hospital
	CapitalCity	Dummy if hospital is located in the capital of each country

DOI: 10.1002/hec

Table II. Descriptive statistics for patient level variables

	Denmark	Finland	Norway	Sweden	All	
Number of observed discharges	15 753 686	12 395 963	11 124 765	18 884 433	58 158 847	
Variable	Mean	Mean	Mean	Mean	Mean	Std. Dev
Quality indicators						
Readm30_Emergency[a]	4.76 %	5.52 %	6.96 %		5.62 %	23.04 %
Readm30_Inpatient	4.95 %	12.67 %	13.84 %	9.99 %	9.93 %	29.91 %
Mort30_LastAdmittance	0.44 %	0.34 %	0.41 %	0.51 %	0.43 %	6.58 %
Mort90_LastAdmittance	0.54 %	0.43 %	0.53 %	0.68 %	0.56 %	7.47 %
Mort180_LastAdmittance	0.61 %	0.46 %	0.62 %	0.79 %	0.64 %	7.96 %
Mort365_LastAdmittance	0.72 %	0.49 %	0.74 %	0.96 %	0.75 %	8.66 %
PSI12_vt_pe	0.123 %	0.053 %	0.090 %	0.104 %	0.096 %	3.119 %
PSI13_Sepsis	0.076 %	0.044 %	0.078 %	0.077 %	0.070 %	2.667 %
PSI15_AccidCutPunc	0.005 %	0.005 %	0.024 %	0.014 %	0.012 %	1.083 %
PSI18_ObstTrauma	0.028 %	0.007 %	0.021 %	0.035 %	0.024 %	1.558 %
BedSores	0.015 %	0.005 %	0.031 %	0.028 %	0.020 %	1.434 %
Patient characteristics						
Male	43.01 %	44.80 %	45.58 %	45.31 %	44.63 %	49.71 %
Agegrp0	2.60 %	1.78 %	2.49 %	1.82 %	2.15 %	14.50 %
Agegrp1_9	4.13 %	6.22 %	6.11 %	6.72 %	5.80 %	23.37 %
Agegrp10_19	6.18 %	6.15 %	6.41 %	7.50 %	6.65 %	24.91 %
Agegrp20_29	9.34 %	8.72 %	9.35 %	8.79 %	9.03 %	28.66 %
Agegrp30_39	12.20 %	10.13 %	12.33 %	10.34 %	11.18 %	31.51 %
Agegrp40_49	11.23 %	11.21 %	11.27 %	10.14 %	10.88 %	31.14 %
Agegrp50_59	15.14 %	15.92 %	13.54 %	12.23 %	14.06 %	34.76 %
Agegrp60_69	18.36 %	17.32 %	16.30 %	16.73 %	17.21 %	37.75 %
Agegrp70_79	12.97 %	14.48 %	12.76 %	14.46 %	13.73 %	34.42 %
Agegrp80_89	6.84 %	7.29 %	8.25 %	9.80 %	8.17 %	27.39 %
Agegrp90	1.02 %	0.79 %	1.20 %	1.47 %	1.15 %	10.67 %
Treatment variables						
TransInOwnHospital	8.90 %	10.07 %	3.03 %	5.27 %	6.85 %	25.25 %
TransInOtherHospital	1.06 %	0.63 %	0.45 %	0.92 %	0.80 %	8.93 %
TransOutOwnHospital	6.18 %	9.39 %	3.16 %	5.00 %	5.90 %	23.57 %
TransOutOtherHospital	0.51 %	0.80 %	0.78 %	0.83 %	0.73 %	8.49 %
Charlson	0.113	0.047	0.265	0.196	0.155	0.665
NumSecDiagnoses	0.568	0.183	0.478	0.524	0.454	1.032
Length of stay						
LOS	1.568	1.457	1.669	1.678	1.599	3.519
Municipal variables						
Population	122 740	113 970	112 512	142 442	125 312	184 461
Unemployment	3.72	9.35	2.23	6.44	5.52	3.35
SocialAssist	1.38	6.73	2.53	4.52	3.76	2.46
SingleFamilies	10.89	20.33	19.87	21.02	17.91	5.48
Foreign	5.75	2.67	5.62	6.17	5.20	3.08
Hospital-municipal variable						
Traveltime	0.461	0.450	0.788	0.446	0.516	0.830

[a]Sweden lacks information on emergency status, therefore this variable only has 39 274 414 valid observations.

2.2.2. Quality indicators. We have calculated performance measures based on 11 quality indicators. Table I lists and defines the variables used in the analysis, and Table II gives descriptive statistics by country. All the indicators are binary variables at the patient level and are therefore presented as rates at the hospital or country levels.

Unlike planned readmissions, emergency readmissions within 30 days of a hospital discharge (but no sooner than the next day) are commonly viewed as a signal of poor medical quality if proper case-mix adjustment has taken place (Leng *et al.*, 1999). Only inpatients are included in this indicator as coding practice for outpatients

DOI: 10.1002/hec

varies between countries. Although some level of readmissions is unavoidable, an emergency readmission could be a sign that the initial treatment was not adequate or that the discharge was premature. We include emergency readmissions for any reason, because poor quality in the initial treatment (e.g., an operation) could well cause a readmission with another diagnosis (e.g., an infection). Country differences in the readmission rates in Table II are considerable, with Denmark at less than 5% and Norway at almost 7%. In Sweden, the coverage of the variable reflecting whether the admission is acute or planned is bad. As a substitute, we also included an indicator for all readmissions as an inpatient, regardless of emergency status. This is clearly more difficult to interpret as a sign of quality, as planned readmissions may be valid parts of a hospital treatment episode. However, in many cases, it will have a service quality dimension, because going in and out of hospitals is usually not appreciated by patients. Table II reveals that there is substantial variation in inpatient readmission patterns between countries.

Mortality rates[4] are the most widely accepted quality indicators. Even though some of the mortalities are unavoidable, lowering mortality is always an improvement. It has the additional advantage of being coded with little possibility of error. We have included four variants with different time perspectives, death within 30, 90, 180, and 365 days. There is a possibility of a person having several hospital stays within the last days of life, so the differential readmission patterns between countries would influence this indicator if the mortality was attributed fully to all hospital stays. We have therefore calculated a mortality dummy only if the stay is the last in the data before death, in order to attribute the death to this particular admission. In order to calculate 365-day mortality, we have collected patient data for the two years 2008 and 2009, and deaths also for 2010.

Patient safety indices (PSIs) are based on OECD standards using secondary diagnoses (Drösler, 2008). Most of these are applicable only to special patient groups, and Table II reveals very small raw rates, almost all less than a 10th of a percent. These also vary considerably between countries, with the Finnish numbers particularly low. The PSIs are based on secondary diagnoses, and we are aware of large differences in coding practices between countries. Secondary diagnoses are rarely reported in Finland, and the rate of PSIs is closely correlated with the reporting of secondary diagnoses (OECD, 2009). Thus, we cannot determine how much of the variation between countries is due to differences in quality and how much to coding, but within-country comparisons may still be valid.

Several other PSI definitions are available but could not be calculated from the available patient registers. Two more PSIs were initially included but turned out to be so infrequent that case-mix adjustments and hospital differences were meaningless. Numerous other quality indicators have been suggested and discarded, most because data were not available for several countries. In many cases, the data available for these indicators were not reliable. Time from referral to admission ('waiting time') could not be included because the definitions of referral date differed across countries and were not available at all for Sweden. Similarly, the time from admission to first procedure ('lead time within hospital') was not registered in Denmark and Sweden.

2.2.3. Case-mix adjusting variables. For the case-mix adjustment procedure, we have used most of the variables available in the patient registers. Ideally, the adjusting variables should capture characteristics of the patients and their illnesses that possibly influence the outcome, whatever the treatment given by the hospital. The primary risk adjustor used is the DRG formed with the common Nordic grouper. Because the division into the more than 700 DRGs is designed to capture most measurable patient differences that may influence costs, they will also capture many of the aspects that influence the expected values of the quality indicators.

The group of patient characteristics shown in Tables I and II comprises gender and age in 10-year groups, with a special infant group of less than 1 year. For data privacy reasons, the precise age was not available in the pooled cross-country dataset. Although partly endogenous, treatment variables are also allowed to adjust for risk, because these may reflect severity. The variables we coded for describing patient transfers in and out of hospital or department (where patient came from and where they went) do not distinguish between transfer

[4]We use the term 'mortality rates' rather than 'case fatality rates' because the latter are usually defined for a specific medical condition rather than for all hospital admissions.

DOI: 10.1002/hec

to/from home, a (non-hospital) health clinic, or a nursing home as we had to use information available in all four countries. Comorbidity is included both as the number of secondary diagnoses and as the Charlson comorbidity index, which in turn is based on information from secondary diagnoses (Charlson *et al*., 1987). LOS may reflect inefficiency in addition to severity (or even quality). LOS is also an endogenous variable to the hospital.

We have also included some characteristics for the patients' residence municipality in order to capture some socioeconomic differences. These variables are, however, not without challenges. Firstly, they are likely to be dependent between patients in each hospital, because most patients come from a limited number of municipalities in the hospital catchment area. In addition, they may to a large extent capture country effects, because there are marked differences between, for example, unemployment levels following the financial crisis. Finally, we have included travel time between the center of the residence municipality of each patient and the hospital,[5] a variable that previously has shown some explanatory power on hospital costs and that may have some also on quality outcomes (Kalseth *et al*., 2011).

3. METHODS

3.1. Case-mix adjustments

For the case-mix adjusted hospital performance measures, we follow Ash *et al*. (2003) and calculate the observed-to-expected ratio of each quality indicator for each hospital. The expected value, and thus the performance measure, is estimated in each of the six different models $m \in (0, \ldots, 5)$.

Each patient i has an (binary) observable quality indicator, ω_{ihk}, and an expected quality indicator, $\hat{\omega}^m_{ihk}$, subscripted by hospital $h \in (1, \ldots, H)$ and DRG $k \in (1, \ldots, K)$. We suppress an index for which indicator we are studying (see Table I for a list of all quality indicators).

The case-mix adjusted hospital performance measures, P^m_h , are calculated by summing all observed patient outcomes and dividing by the sum of all expected patient outcomes

$$P^m_h = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N_{hk}} \omega_{ihk}}{\sum_{k=1}^{K} \sum_{i=1}^{N_{hk}} \hat{\omega}^m_{ihk}}, \tag{1}$$

where P^m_h is the performance indicator for hospital h in model $m \in (0, \ldots, 5)$ and N_{hk} is the number of patients in DRG k at hospital h. Because all our quality indicators are such that a lower number implies better quality, so will a lower value for the performance measure, P^m_h.

The performance measures P^m_h for $m \in (0, \ldots, 5)$ differ in the way we predict $\hat{\omega}^m_{ihk}$. In our simplest model, $m = 0$, we exploit that each hospital has a different composition of DRGs. The predicted quality indicator for patient i, $\hat{\omega}^0_{ihk}$, is thus just the average value of the quality indicator *within* each DRG for all patients across all hospitals. The predicted outcomes of this model can be written as

$$\hat{\omega}^0_{ihk} = \frac{\sum_{g=1}^{H} \sum_{j=1}^{N_{gk}} \omega_{jgk}}{\sum_{g=1}^{H} N_{gk}}, \tag{2}$$

which is independent of i and h and thus equal for all patients in DRG k.

The predicted quality measure, $\hat{\omega}^m_{ihk}$, can be further improved by conditioning on patient characteristics and municipality-specific variables. Because all our quality indicators in this study are binomial variables, the appropriate method is to estimate the conditional probability by the logit model (Greene, 2000; Hosmer *et al*.,

[5]Traveling times are calculated by Google maps using a STATA procedure from Ozimek and Miles (2011).

DOI: 10.1002/hec

2013). However, given the large number of observations, we need not assume that the explanatory variables have the same impact in all DRGs. Rather, *for each* DRG k, we calculate the expected value as the predicted value based on the maximum likelihood estimation of

$$\omega_{ihk}^{m} = \frac{e^{\beta_{0k}^{m} + \boldsymbol{\beta}_{k}^{m} \mathbf{z}_{ihk}^{m} + \varepsilon_{ihk}^{m}}}{1 - e^{\beta_{0k}^{m} + \boldsymbol{\beta}_{k}^{m} \mathbf{z}_{ihk}^{m} + \varepsilon_{ihk}^{m}}}, \tag{3}$$

where ω_{ihk}^{m} is the quality measure for patient i in DRG k at hospital h; the coefficient vectors β_{0k}^{m}, $\boldsymbol{\beta}_{k}^{m}$ are specific to each DRG k and model $m \in (1 \ldots 5)$; $\mathbf{z}_{ihk}^{m}$ is a vector of individual case-mix adjusting variables; and ε_{ihk}^{m} is the error term, which is assumed to be normally distributed.

For each of the K DRGs and 11 indicators, we estimate five different models m, where higher-order models include more explanatory variables z (confer Table I for all case-mix adjusting variables). In model 1, the explanatory variables captured by z are the patient characteristics; in model 2, the vector includes both patient characteristics and the treatment variables; in model 3, the LOS is also added; model 4 includes also municipal characteristics for the patients' resident municipality; while model 5 adds the traveling time between the resident municipality and the treating hospital. The patients' predicted quality measures, $\hat{\omega}_{ihk}^{m}$, are calculated by setting the error term in Equation (3) to zero.

Following Moger and Peltola (2014), there are no hospital dummies in the estimation of Equation (3). Rather, the individual predicted values, $\hat{\omega}_{ihk}^{m}$, are inserted into (1) to calculate a hospital-specific performance index P_{h}^{m} for each quality indicator.[6]

Because our model is an aggregation of the estimates of a large number of non-linear equations, there are no obvious measures of model performance or goodness of fit. The extremely large sample size precludes the use of the Hosmer–Lemeshow (Hosmer and Lemeshow, 1980; Hosmer *et al*., 2013) test of goodness of fit. Instead, we use the Osius and Rojek (1992) normalization of the Pearson chi-squared statistic as outlined in Hosmer *et al*. (2013, p. 164). Following Greene (2000), we also calculate the R-squared based on the sum of the squared errors $\sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{i=1}^{N_{hk}} \left(\omega_{ihk} - \hat{\omega}_{ihk}^{m}\right)^2$ to indicate the share of explained variance. Finally, we calculate and report the area under the curve (AUC) from receiver operating characteristic analysis. The AUC is commonly used for evaluating the ability of predictions from a logistic regression model in discriminating between outcomes and can be interpreted as the probability that, for example, the fatality prediction for a randomly selected patient who died is greater than the fatality prediction for a surviving patient.

3.2. Productivity estimates

The productivity estimates for the hospitals are based on Farrell (1957) who defined (the input oriented) *technical efficiency* as

$$E = \text{Min}\{\theta | (\theta \mathbf{x}_i, \mathbf{y}_i) \in T\}, \tag{4}$$

where $(\mathbf{x}_i, \mathbf{y}_i)$ is the input/output vector for an observation i and T is the technology or production possibility, usually assumed to be specific to year and country. For an input/output vector $(\mathbf{x}, \mathbf{y})$ to be part of the production possibility set, we need to be able to produce $\mathbf{y}$ using $\mathbf{x}$. Efficiency is then the minimal proportionality factor θ on inputs that is consistent with feasibility, that is, what proportion of inputs are necessary to produce the given output vector.

Estimates of efficiency rely on estimates for the specific technology T; but for comparing productivities, only a common reference surface for observations is needed. The literature on Malmquist productivity indices

[6]Had we estimated a single logit, we could have included hospital fixed effects and estimated the performance measure in the first stage. Our setup with DRG-specific coefficients on the z-vector is equivalent to a single estimation with a full set of interaction terms. With 783 DRGs and up to 25 covariates, this would mean simultaneous estimation of up to 20,000 coefficients. Unfortunately, we have not had the necessary programs or machine power available.

DOI: 10.1002/hec

uses an (homogenous in inputs and outputs) envelopment of the technology to estimate changes in technical productivity over time (Førsund and Hjalmarsson, 1987; Grifell-Tatjé and Lovell, 1995). To compare the productivity of two or more observations, we do not need to estimate the separate country-specific and time-specific technologies but may instead rely on an estimate of the meta-frontier or the envelopment of the underlying technologies (Asmild and Tam, 2007). Productivity estimates of individual observations are then compared with this global measure of the highest attainable productivity. Here, we will estimate productivity of a hospital by calculating the reference set T from the pooled set of all hospitals across the Nordic countries and the two years 2008 and 2009 and then comparing individual hospitals to the reference set.

The estimates of the reference set, and therefore of the productivity of each hospital observation, are made using the homogenous version of the non-parametric DEA, one of the two main methods in the productivity literature (Coelli *et al*., 2005; Fried *et al*., 2008). This does not imply an assumption of constant returns to scale technology, because the reference frontier is only a homogenous envelopment of the underlying technology.[7] Because DEA estimates are known to be biased and the statistical properties are not available in closed form, bias-corrected estimates and confidence intervals have been calculated using the bootstrapping algorithm from Simar and Wilson (1998). The average cost per DRG point, which does not use a frontier method, is also calculated.

3.3. Productivity–quality trade-off

This article offers no full behavioral model of the relationship between productivity and quality. We start by calculating the hospital level pairwise Pearson correlation coefficient for each performance indicator, average costs (operating costs/DRG points), and productivity estimates. Additionally, we estimate a simple regression model with random hospital effects, assuming that the unobserved hospital heterogeneity is uncorrelated with the included variables (Greene, 2000)[8]

$$\hat{TP}_{hct} = \boldsymbol{\gamma}_0 + \hat{\mathbf{P}}_{ht}\boldsymbol{\gamma}_1 + \mathbf{x}_{hct}\boldsymbol{\gamma}_2 + \lambda_c + \phi_{hct}, \tag{5}$$

where $\hat{TP}_{hct}$ is the DEA bootstrapped estimate of productivity for hospital h in country c in year t and $\hat{\mathbf{P}}_{ht}$ is a vector of the performance indicators estimated in model 2 (Table V). Note that for estimating the productivity–quality trade-off, we have calculated two performance measures for each hospital, one for each year, instead of pooling both years in a single hospital performance measure. $\mathbf{x}_{hct}$ is a vector of hospital-specific variables, including municipal variables averaged at the hospital level (Table VI); and λ_c contains country fixed effects.

Equation (5) is also estimated for each country separately thus leaving out the country-specific fixed effects.

4. RESULTS

4.1. Case-mix adjustments

Summarizing the more than 700 case-mix adjustment logit regressions for each of the 11 quality indicators is not straightforward. The R-squared statistics (shown in Table A.I of the Appendix) is rather low for all models. This is common in logistic regressions as the outcomes are 0 or 1, while the predictions are almost always between. Because of the large number of observations, R-squared for all models are significantly larger than zero; and for every quality indicator, adding a block of variables significantly increases the share of explained variance.[9] This test therefore gives no direct guidance on the model specification. We note, however, that the

[7] In Kittelsen *et al*. (2015), the productivity estimates are decomposed into scale efficiency, technical efficiency, and country-specific factors using variable returns to scale assumption on the underlying technology in each country. The results there are not sensitive to the use of DEA or the competing stochastic frontier analysis method.

[8] Simultaneous multi-level modeling of the case-mix adjustment is precluded by the computational intractability of the large number of coefficients, confer footnote 6. The large number (58 million) of patient observations would not in itself be a barrier, even though all calculations take much time.

[9] This is equivalent to the adjusted R-squared increasing as we move to larger models. In fact, the numbers for R-squared and adjusted R-squared are not distinguishable with the number of decimals reported in the table.

DOI: 10.1002/hec

DRGs in model 0 explain at most 9% of the variance of the quality indicators. Adding patient characteristics as is performed in model 1 does not change this pattern and hardly adds any explanatory power. Adding the treatment variables in model 2 and LOS in model 3 increases R-squared somewhat for readmission rates, mortality rates, and for PSI13 (sepsis). The municipal and travel time variables of models 4 and 5, respectively, only slightly increase R-squared.

The normalized Pearson goodness-of-fit test (shown in Table A.II of the Appendix) fails to reject the large majority of models. Model 1 is rejected for some of the quality indices and for PSI18 (Obstetric trauma) also, model 2 is rejected; but here, it seems that the problem is that PSI18 only applies to women in specific DRGs and age groups.

Table III. Area under the curve (AUC) based on a 0.1% sample of discharges, using predictions from the full-sample DRG-specific case-mix adjustment regression models with quality indicators as dependent variables.

Model	0	1	2	3	4	5
Cummulative included independent variables	DRGs	+Patient characteristics	+Treatment variables	+Length of stay	+Municipal variables	+Travel time
Dependent variable						
Readm30_Emergency	0.72***	0.73*	0.75***	0.75	0.77***	0.77
Readm30_Inpatient	0.71***	0.73***	0.74***	0.75	0.78***	0.78
Mort30_LastAdmittance	0.92***	0.95***	0.96	0.96	0.96	0.96
Mort90_LastAdmittance	0.91***	0.94***	0.95*	0.95	0.96	0.96
Mort180_LastAdmittance	0.89***	0.94***	0.95*	0.95	0.95	0.95
Mort365_LastAdmittance	0.86***	0.92***	0.93*	0.93	0.94	0.94
PSI12_vt_pe	0.86***	0.88	0.95***	0.95	0.95	0.95
PSI13_Sepsis	0.95***	0.96	0.98***	0.98	0.99	0.99
PSI15_AccidCutPunc	0.95***	0.97	0.99***	0.99	0.99	0.99
PSI18_ObstTrauma	0.98***	0.99	1.00***	1.00	1.00	1.00
BedSores	0.97***	0.98	0.99***	0.99	0.99	0.99

Model *m* includes all variables from model *m-1*. AUC estimates for model *m* that are significantly higher than that of model *m-1* are marked at *0.10; ** 0.05; ***0.01 level. The AUC and the corresponding confidence intervals are estimated using the roctab procedure in Stata 13. In the full sample there are 58 158 847 observations except for Readm30_Emergency which is not registered in Sweden and therefore has only 39 274 414 observations.

More interesting are probably the AUC results shown in Table III.[10] The ability to discriminate between outcomes is very high for all mortality and PSI indicators. For the mortality indicators, the inclusion of patient characteristics significantly increases the AUC estimates and weakly so does the inclusion of treatment variables. For the PSIs, patient characteristics do not seem to matter but treatment variables do. LOS, municipal variables, and travel time do not contribute for these quality indicators. The readmission variables have a clearly different pattern, with lower but considerable AUCs in all models. Here, the inclusion of patient characteristics and treatment variables is significant, as well as the municipal variables. It must be noted, however, that there are large country differences in some of the municipal variables, for example, the number of foreign citizens and the unemployment rates in the wake of the financial crisis.

The statistical evidence seems to favor model 2, with some exceptions. The purpose of these models is to level the field in country and hospital comparisons. The choice of case-mix adjustment model specification must therefore also take account of the problems of country effects in the municipal variables. In addition, the LOS is to a large extent an endogenous variable for the hospital in question and may be more of a mediating than confounding variable. In the further analysis, we therefore use model 2, that is, the model without LOS, the municipal variables, and travel time, returning to these only in the hospital trade-off regressions.

[10] STATA 13 was not able to calculate AUC based on the extremely large samples, so we report AUC results for a 0.1% random subsample stratified on hospitals with 58,159 patients (39,274 patients for Readm30_Emergency).

DOI: 10.1002/hec

Table IV. Country means of case-mix adjusted performance measures (model 2) with 99 % confidence intervals.

	Denmark	Finland	Norway	Sweden
Readm30_Emergency	0.891	1.031	1.103	-
	(0.888 - 0.893)	(1.028 - 1.034)	(1.099 - 1.106)	-
Readm30_Inpatient	0.573	1.235	1.256	0.986
	(0.572 - 0.575)	(1.232 - 1.237)	(1.253 - 1.258)	(0.984 - 0.988)
Mort30_LastAdmittance	0.927	1.037	0.751	1.011
	(0.918 - 0.936)	(1.024 - 1.050)	(0.741 - 0.760)	(1.002 - 1.019)
Mort90_LastAdmittance	0.909	1.043	0.785	1.052
	(0.901 - 0.917)	(1.031 - 1.055)	(0.776 - 0.794)	(1.044 - 1.060)
Mort180_LastAdmittance	0.907	0.989	0.808	1.071
	(0.900 - 0.915)	(0.978 - 0.999)	(0.800 - 0.817)	(1.064 - 1.079)
Mort365_LastAdmittance	0.918	0.877	0.840	1.101
	(0.911 - 0.925)	(0.868 - 0.886)	(0.832 - 0.848)	(1.095 - 1.108)
PSI12_vt_pe	1.153	0.870	0.763	0.992
	(1.131 - 1.174)	(0.842 - 0.898)	(0.742 - 0.783)	(0.974 - 1.011)
PSI13_Sepsis	1.319	1.081	0.718	0.967
	(1.288 - 1.350)	(1.043 - 1.119)	(0.698 - 0.739)	(0.946 - 0.988)
PSI15_AccidCutPunc	0.459	0.681	1.145	0.934
	(0.418 - 0.500)	(0.615 - 0.749)	(1.085 - 1.205)	(0.886 - 0.983)
PSI18_ObstTrauma	0.917	0.393	0.727	1.529
	(0.881 - 0.953)	(0.358 - 0.429)	(0.687 - 0.768)	(1.480 - 1.579)
BedSores	0.752	0.433	1.015	0.992
	(0.713 - 0.791)	(0.389 - 0.478)	(0.969 - 1.062)	(0.956 - 1.027)

4.2. Country and hospital differences

The case-mix adjustments (in model 2) change the relative performance of the countries to some extent. Table IV gives the mean performance measure at the country level, with a 99% confidence interval calculated from the individual patients' predicted values. By construction, each performance measure has a mean of 1.0 when averaging over all four Nordic countries, rendering almost all country-specific performance measures significantly different from the Nordic mean. As the quality measures used are by definition 'measures of low quality', lower performance measures denote higher quality.

The quality measures in Table IV do not give a uniform picture of the quality of care in any of the Nordic countries. Neither do they indicate any clear ranking of the countries. While Denmark has clearly fewer emergency readmissions, Norway has the lowest mortality rates. The inpatient readmission rates, on the other hand, are higher in Norway and Finland than in Denmark and Sweden. PSI12 (pulmonary/deep vein thrombosis) and PSI13 (sepsis) are the lowest for Norway; PSI15 (accidental cut, puncture, or haemorrhage during medical care) is the lowest in Denmark; while Finland has the lowest score for PSI18 (obstetric trauma) and bed sores.

Table V. Hospital differences in case-mix adjusted performance measures (model 2)

	Share of hospitals with performance measure significantly different from 1 at 95% level					ANOVA
	Denmark	Finland	Norway	Sweden	Total	F
Readm30_Emergency	89 %	91 %	81 %		86 %	774.4***
Readm30_Inpatient	100 %	97 %	87 %	87 %	91 %	3635.3***
Mort30_LastAdmittance	68 %	59 %	85 %	64 %	70 %	82.6***
Mort90_LastAdmittance	79 %	66 %	81 %	85 %	79 %	96.6***
Mort180_LastAdmittance	75 %	53 %	72 %	85 %	73 %	98.9***
Mort365_LastAdmittance	86 %	66 %	74 %	81 %	77 %	112.2***
PSI12_vt_pe	75 %	53 %	74 %	51 %	63 %	63.9***
PSI13_Sepsis	57 %	41 %	74 %	62 %	61 %	55.4***
PSI15_AccidCutPunc	85 %	34 %	30 %	32 %	41 %	13.3***
PSI18_ObstTrauma	57 %	81 %	48 %	62 %	60 %	36.2***
BedSores	54 %	80 %	39 %	43 %	51 %	19.7***
Number of hospitals	28	32	47	53	160	160

ANOVA tests for differences in hospital performance and the significance of the F-values are marked at *0.10; ** 0.05; ***0.01 level.

DOI: 10.1002/hec

Hospital differences are difficult to summarize, but Table V shows the percentage of hospitals across both years with performance measures significantly different from the Nordic mean of 1. This holds for almost all of the readmission variables, and for a large majority of the mortality rates, but to a lesser and mixed extent for the PSIs. For readmission variables, Denmark and Finland have the largest shares of hospitals with performance measures different from the Nordic mean. Sweden has the largest share of hospitals with significantly different means in two mortality measures. For mortality within 30 days, Norway has the largest share of hospitals with significantly different means. The last column of Table V shows the significance of the hospitals in explaining the variation remaining after the case-mix adjustment of model 2, based on a linear ANOVA test of the difference between observed and predicted values (Greene, 2000). The results show that hospitals are significantly different from each other in their performance measures for all quality indicators. Given the very large number of patient observations, the *F*-values are not particularly high for the mortality indicators and definitely weak for the PSIs.

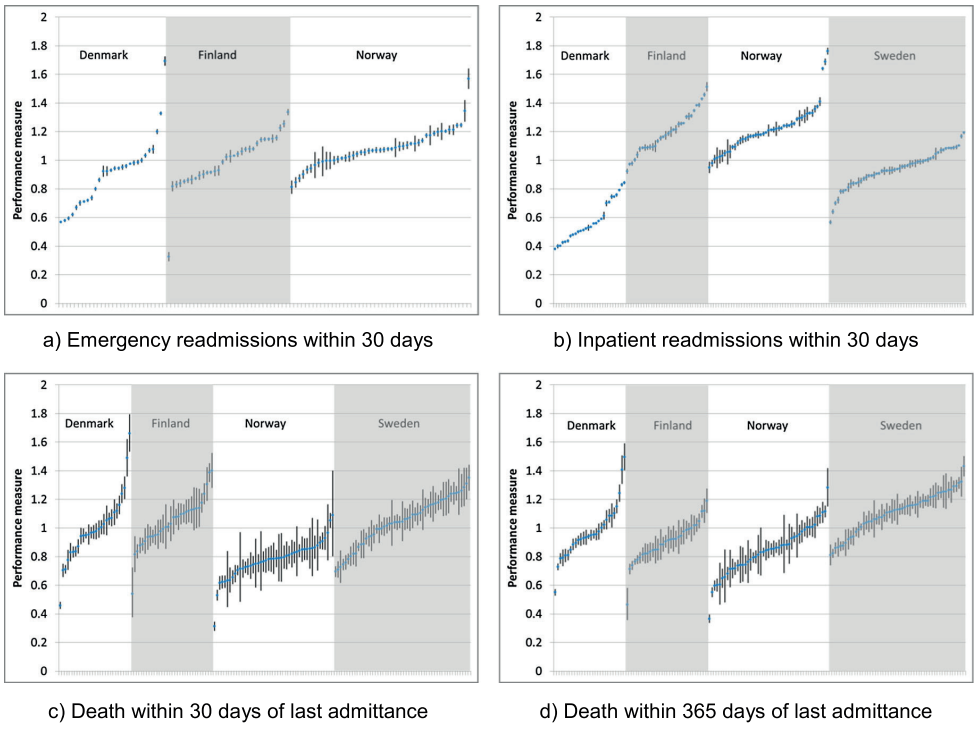


Figure 1. Selected case-mix adjusted performance measures for hospitals sorted by country, with 99% confidence intervals. Lower numbers indicate better quality.

Figure 1 plots four of the performance measures and their 99% confidence intervals for the individual hospitals sorted by countries. For emergency readmissions (panel a), the confidence intervals are very narrow, which means that there are significant differences between most hospitals. There is mostly a clear ranking of hospitals within countries, because each hospital performance measure is mainly outside the range of other hospitals' confidence intervals. As noted, Denmark has the lowest emergency readmission rates, but there is

DOI: 10.1002/hec

some overlap with the Finnish and Norwegian hospitals. It was not possible to compile this indicator for Sweden. Inpatient readmissions (panel b) show even greater differences, with all Danish hospitals having significantly lower rates than all Finnish and Norwegian hospitals. The rates of Swedish hospitals fall mostly between Danish and both Finnish and Norwegian hospitals.

For 30-day or 365-day mortality, the confidence intervals are wider, but most hospitals are still significantly different from the mean and from each other. Most Norwegian hospitals have significantly lower 30-day mortality than hospitals in the other countries, but these differences are less marked when comparing 365-day mortality (panels c and d, respectively).

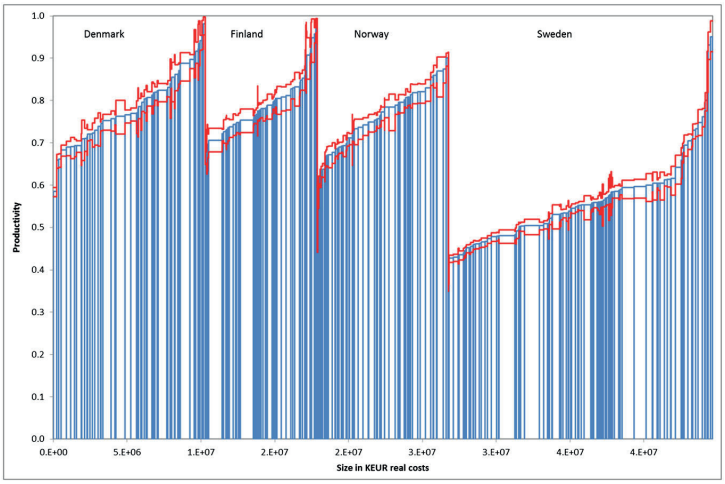


Figure 2. Salter diagram of bootstrapped DEA hospital productivity estimates sorted by country with 95% confidence intervals. The width of each column is proportional to hospital size measured by real costs.

4.3. Productivity

To look into the possible trade-off between hospital productivity and quality, we first had to estimate hospital productivity. As noted in 2.2., we here use a common Nordic version of the NordDRG grouper, which makes it possible to compile hospital output measures that are comparable between countries. Figure 2 shows the bias-corrected DEA productivity estimates of the hospitals sorted by country and productivity levels, with the width of the bars proportionate to hospital costs. Bootstrapped 95% confidence intervals are also shown.

The figure confirms the previous results that Finnish hospitals are on average more productive than in the other Nordic countries, even though Denmark is almost as productive (Medin *et al*., 2011; Kittelsen *et al*., 2008; Linna *et al*., 2010; Kittelsen *et al*., 2015). Even Norway has not much of a cost disadvantage in this analysis, a clear catching up from previous studies. Sweden, however, still lags behind. As a first robustness test, average costs (real costs per DRG point) have also been calculated and show essentially the same picture with a correlation of −88.6%. We have also recalculated the DRG points using calibrated Swedish DRG weights and results are again very similar, with a correlation between productivity estimates of 90.2%. Table A.III in the Appendix shows the mean hospital inputs, outputs, and productivity estimates for each country.

DOI: 10.1002/hec

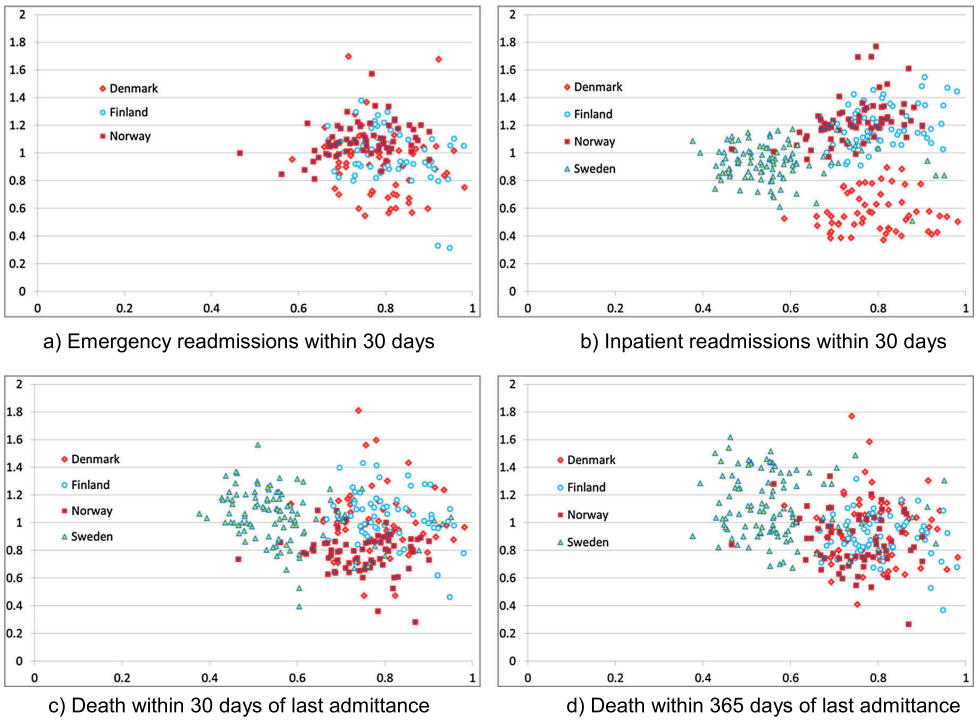


Figure 3. Selected case-mix adjusted performance measures for hospitals (vertical) plotted against estimated productivity (horizontal). Better joint performance is low performance measure and high productivity (lower right).

4.4. Productivity–quality trade-off

When productivity estimates are plotted against four of the performance measures in Figure 3, one finds no strong correlations. In all panels, the optimal frontier would be at the lower right with the highest productivity and the lowest performance measure, keeping in mind that a low performance measure indicates higher quality. In panel a, there is a slightly negative correlation ($r = -0.155$) between productivity and emergency readmissions, implying no trade-off between high quality and high productivity. There is a slight tendency for low emergency readmission rates to go together with high productivity in Finland, but the main impression is a large dispersion. Panel b shows some positive correlation, implying that having high productivity goes together with high number of inpatient readmissions. There seems to be a trade-off between quality and productivity but only in so far as the inpatient readmission rate is a valid quality indicator.

For the two mortality rates shown in Figure 3, mortality within 30 days and mortality within 365 days of hospital episode, there is a clear negative correlation between productivity and performance measures, which is strongest for 365-day mortality. This would imply that there is no trade-off between productivity and quality, and it is possible to improve both productivity and quality at the same time.

The pairwise correlations between measures of productivity and quality in Figure 3 are reported in the first column of Table VI, which draws on the full correlation matrix in Table A.IV in the Appendix.

DOI: 10.1002/hec

Table VI. Productivity-performance correlations and trade-off regressions (GLS random hospital effects models)

	Pairwise correlations with estimated productivity	Linear regression on dependent variable estimated productivity					
		I. All countries	II. Without Sweden	III. Denmark	IV. Finland	V. Norway	VI. Sweden
Constant		0.938 ***	0.951 ***	1.311 ***	0.852 ***	0.688 ***	0.748 ***
		(0.092)	(0.102)	(0.328)	(0.159)	(0.157)	(0.192)
Performance measures							
Readm30_Emergency	−0.155 ***		−0.002	0.110	−0.143 **	0.042	
			(0.038)	(0.082)	(0.066)	(0.065)	
Readm30_Inpatient	0.121**	0.067 *	0.130 ***	0.023	0.196 ***	0.148 **	−0.088
		(0.040)	(0.043)	(0.120)	(0.076)	(0.072)	(0.097)
Mort30_LastDischarge	−0.233***	−0.110***	−0.149 ***	−0.307***	−0.066	−0.092	−0.040
		(0.028)	(0.036)	(0.076)	(0.058)	(0.086)	(0.046)
Hospital variables							
Number of patients	0.095	−1.73E-08	−3.48E-08	−1.66E-07	−1.36E-07	2.06E-07*	−1.17E-07
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
UniversityHospital	0.130 **	0.013	0.020	0.012	−0.006	−0.034	0.095
		(0.023)	(0.024)	(0.102)	(0.058)	(0.040)	(0.061)
CapitalCity	0.102 *	−0.009	−0.008	0.108	−0.085	−0.027	0.151
		(0.040)	(0.044)	(0.152)	(0.207)	(0.081)	(0.087)
Hospital average of municipal variables							
Population	0.058	2.01E-08	3.38E-09	8.64E-08	9.64E-07	3.58E-08	1.92E-09
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Unemployment	−0.122 **	0.002	−0.001	0.003	−0.003	−0.013	0.004
		(0.002)	(0.003)	(0.008)	(0.004)	(0.011)	(0.004)
SocialAssist	−0.112 *	−0.003	−0.002	0.009	0.012	−0.030	−0.005
		(0.009)	(0.012)	(0.078)	(0.014)	(0.042)	(0.014)
SingleFamilies	−0.360***	−0.004	−0.004	−0.019	−0.002	0.003	−0.004
		(0.004)	(0.004)	(0.028)	(0.007)	(0.007)	(0.008)
Foreign	−0.167***	0.000	−0.010	−0.031	−0.039 **	−0.005	0.006
		(0.006)	(0.007)	(0.029)	(0.019)	(0.010)	(0.010)
Traveltime	−0.079	−0.064***	−0.073 ***	0.094	−0.057	−0.055 **	−0.057
		(0.025)	(0.023)	(0.207)	(0.092)	(0.025)	(0.087)
Country dummies							
Denmark		−0.014	0.056				
		(0.062)	(0.067)				
Norway		−0.069	−0.057				
		(0.044)	(0.047)				
Sweden		−0.214***					
		(0.035)					
R-squared							
Within		0.054	0.010	0.023	0.058	0.060	0.036
Between		0.003	0.046	0.033	0.017	0.042	0.097
Overall		0.012	0.081	0.063	0.074	0.096	0.079
Number of observations		292	186	56	64	66	106

Standard errors in (). Significant coefficients are marked at * 0.10; ** 0.05; *** 0.01 level. Hausman tests reject random effects for model I (X^2=25.7), but accepts random effects for model II (X^2=11.9).

DOI: 10.1002/hec

The next columns of Table VI show the results of the random effects regressions on the bias-corrected productivity estimate with the main performance measures at the hospital level as explanatory variables (Equation (5)). For reasons of collinearity, only one mortality measure, mortality within 30 days, is included. The results for the PSIs are also deemed too weak to be valid as measures of quality for hospitals as a whole and therefore not included as explanatory variables. Column I excludes the emergency readmissions, because Sweden has no observations on this measure, while column II instead excludes the Swedish observations. The last four columns show regressions for each country separately. The country-specific models have generally less explanatory power due to the low number of observations.

The negative pairwise correlation between productivity and emergency readmissions disappears when modeled simultaneously with the other performance measures and the control variables. Looking at individual countries, only in Finland is there a significant negative coefficient, indicating no trade-off between productivity and emergency readmission rates. In Denmark and Norway, the coefficients are positive but insignificant. The lack of significance in some of the country-specific associations may partly reflect the low number of observations. For inpatient readmission, the positive correlation, indicating a trade-off, carries through to the regression coefficients, with high inpatient readmission rates being associated with high productivity, particularly in Finland and Norway.

As in Figure 3, there seems to be some association of high quality and high productivity when using 30-day mortality as a quality measure on a pooled dataset with all countries. However, in the country-specific regressions, the association is only significant for Denmark.

Of the controls, Sweden has significantly lower productivity, while the Norway and Denmark dummies are not significant. In common with previous studies (Kalseth *et al*., 2011), one finds a negative association between productivity and the traveling time, which seems to be influenced mainly by the Norwegian observations. No other controls are significant.

5. DISCUSSION

Our study confirms the productivity differences from previous studies with a similar ranking of the countries where Finnish hospitals have the highest productivity estimates followed by Denmark and Norway and last Sweden. Overall, there is a weak pattern that Norway and Denmark show higher performance in quality and Sweden lower performance; thus, there is no trade-off between productivity and quality at the country level. The distance between hospitals in Sweden and the other countries is even larger when taking quality aspects into account, although the confidence intervals for several indicators are overlapping, which makes ranking of hospitals within each country difficult.

Case-mix adjustment of quality indicators is important; and in some cases, the ranking of countries changed as a result of the adjustment. Our result shows that there are major and significant differences in the included case-mix adjusted quality indicators, especially across countries. One example is inpatient readmission within 30 days where Denmark has less than half compared with the other countries. Also, mortality within 30 days after last admittance varies considerably with Norway having the lowest rates and Finland the highest. The ranking of the countries changes for the longer periods up to 365 days after admission, where Finland has lower mortality rates than both Sweden and Denmark. The case-mix adjusted performance measures for hospitals show larger differences for readmissions than for mortality after last admission. The smaller variation for mortality is in line with other studies about mortality differences. In Sweden and Norway, 30-day mortality of AMI was lower than that in Finland; but for stroke and hip fracture, there was no difference between the three countries (Heijink *et al*., 2015; Häkkinen *et al*., 2015; Hagen *et al*., 2015; Medin *et al*., 2015; Peltola *et al*., 2015). Other quality indicators largely confirm these country level differences and are based on a low frequency of events, and some PSIs are not usable for

DOI: 10.1002/hec

quality comparisons at the country level because of poor or divergent coding. PSIs may still reveal important differences between and within countries with lower general standard of health care, such as in some developing countries. For most quality indicators, the performance measures reveal room for improvement in Nordic hospitals.

The hospital variables do not contribute to the explanation of the differences in productivity. The number of patients, as a proxy for size, is not significant nor is university hospitals or capital city. A number of municipal variables were tested, and only travel time was associated with higher costs and then only in Norway.

Regarding the cost–quality trade-off, there is a statistical significant negative relationship between hospital-level productivity and mortality within 30 days. Assuming that the case-mix adjustment is adequate, the driving mechanism seems to be that treating dying patients is costly for the hospitals, and the efficient hospitals are those that are better at preventing mortality. This relationship is valid for all countries, but in the analysis for each country, it is only significant for Denmark. Hospitals with higher inpatient readmissions within 30 days have a tendency for also having higher productivity. This relationship is significant for Finland and Norway indicating a productivity–quality trade-off.

The conclusion from the review by Hussey *et al*. (2013) was that evidence of the direct association between productivity and quality is inconsistent but that the association is small to moderate. Of these, six studies used similar outcome indicators as in our study (Barnato *et al*., 2010; Bradbury *et al*., 1994; Carey and Burgess, 1999; Deily and McKay 2006; McKay and Deily, 2008; Fleming, 1991). Among these studies, three indicate a clear negative relationship. Compared with our study, these studies include several limitations. Outcome is often measured using in-hospital mortality, or data are restricted to Medicare patients (over 65). In addition, the cost measures in the US studies usually exclude the costs for the physician, whereas in our study, the wages of physicians are included.

In some of the recent studies referenced in the introduction, the outcome measures are similar to the present study. Stukel *et al*. (2012) found that higher hospital spending intensity was associated with better survival and lower admission rates. Stargardt *et al*. (2014) confirmed the trade-off between costs and outcomes, estimating that an increase of costs by €100 leads to a reduction in mortality risk by 0.4%. Doyle *et al*. (2015) also found that patients brought to a higher cost hospital have lower mortality. These studies show a different result to ours by finding a productivity–quality trade-off. The US study at hospital level by Birkmeyer *et al*. (2012) came to a somewhat different conclusion. The study found a strong positive correlation between complication rates and episode payments, indicating that efforts to improve surgical quality may reduce costs. This study differed in the cost estimation as the time window was extended to 30 days after discharge, which could explain a different correlation pattern where low quality leads to high costs after discharge.

Controlling sufficiently for patient–case mix is a major concern and may be a limitation of our study. The most important variables included are the DRGs formed with the common Nordic grouper, age, gender, hospital transfers, and comorbidities. Still, as coding practices differ across the countries (especially in the reporting of secondary diagnoses), true differences in risk factors at the patient level may not be sufficiently captured. There is a need for improvement in harmonization in coding across countries.

The use of average cost to weigh the patient discharges in different DRGs does not necessarily reflect the social willingness to pay for the different treatment groups. Basing these weights on average costs in two Finnish municipalities poses additional problems if these weights then reflect costs or incentives that are particular for Finland. However, using calibrated Swedish weights showed results to be quite robust, and previous studies that exploit the variation in the use of activity-based financing in Nordic hospitals have found little effect on productivity (Kittelsen *et al*., 2008). Still, there is a clear need for further research on the effect of different weight sets for the measurement of productivity.

DOI: 10.1002/hec

Another limitation of this study is that health outcomes and productivity measures are reported at the hospital level, which may conceal differences in outcomes across medical conditions. In a recent European study (Häkkinen *et al*., 2015), the results indicated that there was no correlation either at the national or at the hospital level, between the quality in treatment of two different acute conditions (stroke and AMI). The results indicate that the quality of treatment for one specific health problem (disease) cannot be used as a proxy for hospital level overall quality of care.

Finland treats patients with incurable diseases to a larger extent outside acute hospitals, which may have an impact on hospital mortality figures. Still this does not explain the differences in-hospital mortality between Norway and Sweden, and some of these effects may have been controlled for by the use of age groups in the case-mix adjustments.

Finally, it is an important limitation that causal inferences cannot be drawn from the regressions; instead, these indicate the strength of statistical associations. In particular, the productivity–quality trade-off regressions are not set in a behavioral model with adequately specified econometric structure.

6. CONCLUSION

The results show that there are significant differences between countries on most measured quality indicators. Case-mix adjustments are necessary but explain only a minor portion of quality variation. There are significant differences also between hospitals within countries, but only the readmission and mortality measures show enough differences to rank the majority of hospitals. For PSIs, the confidence intervals overlap too much for rankings to be meaningful. The PSI events are too infrequent in the Nordic countries to discriminate between chance and true hospital or country differences, and are generally prone to be invalid for country comparisons due to differences in coding practices. This highlights the need for continuous improvements in the harmonization of coding systems and patient registry information. At this point in time, only some patient registry-based quality indicators are useable for international comparisons, especially if one looks beyond the Nordic countries. If individual hospital managers are to learn from other hospitals, and national policy makers are to learn from other countries, comparable data must be provided. This does not necessarily imply use of common DRG systems or incentives but that the underlying diagnosis, procedure, and case-mix adjustment codes have the same content.

While previous findings on the relative productivity of the hospitals in the Nordic countries are confirmed, there is no clear pattern that any country has higher or lower quality on all measures. This may be due to the limitations of the available data as discussed earlier. This may also be due to that the treatment patterns and practices vary a lot between countries, even for countries that are as similar as Denmark, Finland, Norway, and Sweden. This is consistent with previous findings that efficiencies are similar across countries but that there are country-specific factors that make the production possibilities significantly different (Kittelsen *et al*., 2015). Unfortunately, statistical methods have difficulty in identifying the effect of country-specific factors with only four countries. Again, the use of data from countries outside the Nordic region could give a better foundation for general results.

The evidence for a trade-off or a positive association between quality and productivity varies between the different performance measures. There seems to be a trade-off between productivity and better (lower) inpatient readmission rates, but high productivity is associated with lower mortality rates. Policies aimed at reducing readmission rates may be costly. There may be differences between emergency and planned readmissions in this regard, but it is important to look into how incentives for readmissions vary between countries. Policies aimed at decreasing mortality rates may reduce costs and increase productivity at the hospital level. For mortality at least, there seems to be a possibility of improving both quality and productivity.

DOI: 10.1002/hec

APPENDIX I

Table A.I R-squared for case-mix adjustment regressions of quality indicators

Model	0	1	2	3	4	5
Cummulative included independent variables	DRGs	+Patient characte-ristics	+Treatment variables	+Length of stay	+Municipal variables	+Travel time
Dependent variable						
Readm30_Emergency	0.05***	0.05***	0.08***	0.08***	0.09***	0.09***
Readm30_Inpatient	0.08***	0.09***	0.11***	0.11***	0.15***	0.15 ***
Mort30_LastAdmittance	0.09***	0.12***	0.14***	0.15***	0.16***	0.16***
Mort90_LastAdmittance	0.09***	0.12***	0.15***	0.16***	0.17***	0.17***
Mort180_LastAdmittance	0.09***	0.12***	0.15***	0.16***	0.16***	0.16***
Mort365_LastAdmittance	0.08***	0.11***	0.14***	0.14***	0.15***	0.15***
PSI12_vt_pe	0.01***	0.01***	0.01***	0.01***	0.02***	0.02***
PSI13_Sepsis	0.02***	0.03***	0.06***	0.07***	0.08***	0.08***
PSI15_AccidCutPunc	0.01***	0.02***	0.04***	0.04***	0.06***	0.07***
PSI18_ObstTrauma	0.10***	0.10***	0.11***	0.11***	0.11***	0.11***
BedSores	0.00***	0.01***	0.02***	0.02***	0.04***	0.04***

Model *m* includes all variables from model *m-1*. R^2 estimates for model *m* that are significantly higher than that of model *m-1* are marked at * 0.10; ** 0.05; *** 0.01 level. Calculations have been performed using Stata 13 matrix commands. There are 58 158 847 observations except for Readm30_Emergency which is not registered in Sweden and therefore has only 39 274 414 observations.

Table A.II Normalized Pearson chi-squared statistics (Z-values) for case-mix adjustment regressions of quality indicators

Model	0	1	2	3	4	5
Cummulative included independent variables	DRGs	+Patient characte-ristics	+Treatment variables	+Length of stay	+Municipal variables	+Travel time
Dependent variable						
Readm30_Emergency	0.03	2.093E+11***	0.00	0.00	0.00	0.00
Readm30_Inpatient	0.03	234.34***	0.00	0.00	0.00	0.00
Mort30_LastAdmittance	0.03	1.14	0.00	0.00	0.00	0.00
Mort90_LastAdmittance	0.03	1.30	0.00	0.00	0.00	0.00
Mort180_LastAdmittance	0.03	2.27**	0.00	0.00	0.00	0.00
Mort365_LastAdmittance	0.03	12.32***	0.00	0.00	0.00	0.00
PSI12_vt_pe	0.03	1.90*	0.00	0.00	0.00	0.00
PSI13_Sepsis	0.03	1.38	0.00	0.00	0.00	0.00
PSI15_AccidCutPunc	0.03	−2.13**	0.00	0.00	0.00	0.00
PSI18_ObstTrauma	0.03	−22.80***	−13.18***	−0.31	0.00	0.00
BedSores	0.03	−0.71	0.00	0.00	0.00	0.00

[1]

Model *m* includes all variables from model *m-1*. Z-values for model *m* that are significantly different from zero are marked at * 0.10; ** 0.05; *** 0.01 level. Calculations have been performed using Stata 13 matrix commands. There are 58 158 847 observations except for Readm30_Emergency which is not registered in Sweden and therefore has only 39 274 414 observations.

Table A.III Hospital productivity model - Number of observations, mean input and outputs and bootstrapped DEA productivity estimates with 95% confidence intervals

	Denmark	Finland	Norway	Sweden	Total
Observations					
2008	28	32	37	52	149
2009	28	32	29	54	143
Input					
Real costs in KEUR	183 778	118 682	134 550	168 804	152 948
Outputs					
Medical inpatients	126 116	58 333	78 822	69 611	80 057
Surgical inpatients	66 804	56 178	59 928	55 550	58 835
Outpatients	115 661	78 565	64 921	67 999	78 760
Productivity estimates					
Mean	0.791	0.805	0.746	0.566	0.702
95% CI	(0.773 - 0.805)	(0.787 - 0.818)	(0.732 - 0.756)	(0.556 - 0.574)	(0.691 - 0.712)

DOI: 10.1002/hec

Table A.IV Hospital level pairwise Pearson correlation coefficients (first part)

			Performance measures					
	Average Cost (Operating costs/ DRG-points)	Productivity estimate	Readm30_ Emergency	Readm30_ Inpatient	Mort30_ LastDischarge	Mort90_ LastDischarge	Mort180_ LastDischarge	Mort365_ LastDischarge
Average Cost (Operating costs/ DRG-points)	1.00							
Productivity estimate	−0.89*	1.00						
Performance measures								
Readm30_Emergency	0.10	−0.15*	1.00					
Readm30_Inpatient	−0.04	0.12*	0.35*	1.00				
Mort30_LastDischarge	0.13*	−0.23*	0.08	−0.23*	1.00			
Mort90_LastDischarge	0.17*	−0.28*	0.10	−0.18*	0.97*	1.00		
Mort180_LastDischarge	0.24*	−0.34*	0.12*	−0.19*	0.89*	0.96*	1.00	
Mort365_LastDischarge	0.30*	−0.37*	0.11	−0.22*	0.63*	0.75*	0.90*	1.00
Hospital variables								
Number of patients	−0.06	0.09	−0.05	−0.04	−0.14*	−0.17*	−0.16*	−0.12*
Case-Mix Index (CMI)	0.35*	−0.05	−0.05	0.06	−0.19*	−0.19*	−0.16*	−0.06
LOS deviation	0.49*	−0.50*	0.17*	0.10	0.30*	0.33*	0.30*	0.21*
Outpatient share	−0.36*	0.03	−0.10	−0.17*	0.26*	0.24*	0.18*	0.03
UniversityHospital	0.02	0.13*	0.14*	−0.02	−0.28*	−0.31*	−0.31*	−0.23*
CapitalCity	−0.08	0.10	0.09	−0.08	−0.27*	−0.31*	−0.30*	−0.22*
Hospital average of municipal variables								
Population	0.01	0.06	−0.02	−0.01	−0.23*	−0.25*	−0.23*	−0.15*
Unemployment	0.10	−0.12*	0.00	0.10	0.38*	0.44*	0.43*	0.34*
SocialAssist	0.10	−0.11	0.12*	0.42*	0.22*	0.29*	0.25*	0.10
SingleFamilies	0.37*	−0.36*	0.33*	0.58*	−0.06	0.02	0.04	0.04
Foreign	0.18*	−0.17*	0.09	−0.22*	−0.26*	−0.28*	−0.21*	−0.03
Traveltime	0.14*	−0.08	0.04	0.18*	−0.31*	−0.27*	−0.21*	−0.12*

* Significant at the 0.05 level.

CONFLICT OF INTEREST

The authors have no conflict of interest. None of the authors have received grants, speakers fees, etc., from any relevant commercial body within the past 2 years.

ETHICAL STATEMENT

Permission to use patient data from Norwegian Regional Ethics Committee (Ref: 2011/930/REK), from the Norwegian Data Protection Authority (Ref: 11/01210-3/THE), from the Regional ethical review board in Stockholm (Dnr: 2011/213-31/1), and from the Danish Data Protection Agency.

ACKNOWLEDGEMENTS

We acknowledge the contribution of other participants in the Nordic Hospital Comparison Study Group (http://www.thl.fi/nhcsg/) and the EuroHOPE project (http://www.eurohope.info/) in the collection of data and discussion of study design and results. We thank the European Union (7FP grant agreement no. 241721), the Research Council of Norway (grant 214338/H10), as well as our respective employers, for financial contributions, and we also thank the participants of various seminars for helpful comments. We finally thank two anonymous referees for very insightful and useful comments.

REFERENCES

Anthun KS, Goude F, Häkkinen U, Kittelsen SAC, Kruse M, Medin E, Rehnberg C, Rättö H. 2013. Eurohope hospital level analysis: material, methods and indicators. *Eurohope discussion papers* Helsinki, THL.

Ash AS, Shwartz M, Peköz EA. 2003. Comparing outcomes across providers. In *Risk Adjustment for Measuring Health Care Outcomes*, LI Iezzoni.Health Administration Press: Chicago; 297–333.

DOI: 10.1002/hec

Table A.IV Hospital level pairwise Pearson correlation coefficients (second part)

	Hospital variables						Hospital average of municipal variables					
	Number of patients	Case-Mix Index (CMI)	LOS deviation	Outpatient share	University Hospital	Capital City	Population	Unemployment	Social Assist	Single Families	Foreign	Traveltime
Average Cost (Operating costs/ DRG-points)												
Productivity estimate												
Performance measures												
Readm30_Emergency												
Readm30_Inpatient												
Mort30_LastDischarge												
Mort90_LastDischarge												
Mort180_LastDischarge												
Mort365_LastDischarge												
Hospital variables												
Number of patients	1.00											
Case-Mix Index (CMI)	−0.04	1.00										
LOS deviation	−0.29*	0.10	1.00									
Outpatient share	−0.07	−0.77*	−0.02	1.00								
UniversityHospital	0.54*	0.08	−0.16*	−0.22*	1.00							
CapitalCity	0.29*	0.04	−0.24*	−0.08	0.46*	1.00						
Hospital average of municipal variables												
Population	0.36*	0.12*	−0.20*	−0.20*	0.39*	0.77*	1.00					
Unemployment	−0.08	−0.10	0.22*	0.29*	−0.10	−0.21*	−0.18*	1.00				
SocialAssist	−0.03	−0.07	0.27*	0.21*	−0.05	−0.15*	−0.11	0.81*	1.00			
SingleFamilies	−0.07	0.04	0.33*	−0.05	−0.03	0.04	0.10	0.28*	0.57*	1.00		
Foreign	0.22*	0.12*	−0.08	−0.20*	0.39*	0.68*	0.67*	−0.34*	−0.30*	0.06	1.00	
Traveltime	0.00	0.10	−0.02	−0.24*	0.06	−0.16*	−0.21*	−0.20*	−0.11*	0.15*	−0.13*	1.00

Asmild M, Tam F. 2007. Estimating global frontier shifts and global Malmquist indices. *Journal of Productivity Analysis* **27**: 137–148.

Barnato AE, Chang C-CH, Farrell MH, Lave JR, Roberts MS, Angus DC. 2010. Is survival better at hospitals with higher "end-of-life" treatment intensity. *Medical Care* **48**(2): 125–132.

Birkmeyer JD, Gust C, Dimick JB, Birkmeyer NJO, Skinner JS. 2012. Hospital quality and the cost of inpatient surgery in the United States. *Annals of Surgery* **255**(1): 1–5.

Bradbury RC, Golec JH, Steen PM. 1994. Relating hospital health outcomes and resource expenditures. *Inquiry* **31**(1): 56–65.

Carey K, Burgess JF. 1999. On measuring the hospital cost/quality trade-off. *Health Economics* **8**(6): 509–520.

Carey K, Stefos T. 2011. Measuring the cost of hospital adverse patient safety events. *Health Economics* **20**(12): 1417–1430.

Charlson ME, Pompei P, Ales KL, Mackenzie CR. 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases* **40**(5): 373–383.

Coelli T, Rao DSP, O'donnell CJ, Battese GE. 2005. *An Introduction to Efficiency and Productivity Analysis*, Springer Verlag: New York.

Deily ME, McKay NL. 2006. Cost inefficiency and mortality rates in Florida hospitals. *Health Economics* **15**: 419–431.

Donabedian A. 1966. Evaluating the quality of medical care. *The Milbank Memorial Fund Quarterly* **44**(3): 166–206.

Doyle JJJ, Graves JA, Gruber J, Kleiner SA. 2015. Measuring returns to hospital care: evidence from ambulance referral patterns. *Journal of Political Economy* **123**(1): 170–214.

Drösler S. 2008. Facilitating cross-national comparisons of indicators for patient safety at the health-system level in the OECD countries. *OECD HEALTH TECHNICAL PAPERS*. Paris, OECD.

Farrell MJ. 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society* **120**: 253–281.

Fleming ST. 1991. The relationship between quality and cost: pure and simple? *Inquiry* **28**: 29–38.

Førsund FR, Hjalmarsson L. 1987. *Analyses of Industrial Structure: A Putty-clay Approach*, Almqvist & Wiksell International: Stockholm.

Fried HO, Lovell CK, Schmidt SS. 2008. *The Measurement of Productive Efficiency and Productivity Growth*, Oxford University Press: Oxford.

Greene WH. 2000. *Econometric Analysis*, Prentice-Hall: New Jersey.

Grifell-Tatjé E, Lovell CK. 1995. A note on the Malmquist productivity index. *Economics Letters* **47**(2): 169–175.

Gryna FM. 1999. Quality and costs. In *Juran's Quality Handbook*, Juran JM, Godfrey AB (eds.), McGraw-Hill: New York.
Gutacker N, Bojke C, Daidone S, Devlin NJ, Parkin D, Street A. 2013. Truly inefficient or providing better quality of care? Analysing the relationship between risk-adjusted hospital costs and patients' health outcomes. *Health Economics* **22**(8): 931–947.
Hagen T, Häkkinen U, Belicza E, Fattore G, Gaude F. 2015. Acute myocardial infarction, use of percutaneous coronary intervention, and mortality: a comparative effectiveness analysis covering seven European countries. *Health Economics* **24**(Suppl. 2): 88–101.
Häkkinen U, Rosenqvist G, Iversen T, Rehnberg C, Seppälä T, Kohavakka R. 2015. Outcome, cost and their relationship in the treatment of ami, stroke and hip fracture in European hospitals. *Health Economics* **24**(Suppl. 2): 116–139.
Heijink R, Engelfriet P, Rehnberg C, Kittelsen SAC, Häkkinen U. 2015. A window on regional variation in healthcare: insights from EuroHOPE. *Health Economics* **24**(Suppl. 2): 164–177.
Hosmer DW, Lemeshow S. 1980. Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods* **9**(10): 1043–1069.
Hosmer DW, Lemeshow S, Sturdivant RX. 2013. *Applied Logistic Regression*, John Wiley and Sons: Hoboken, New Jersey.
Hussey PS, Wertheimer S, Mehrotra A. 2013. The association between health care quality and cost: a systematic review. *Annals of Internal Medicine* **158**(1): 27–34.
Hvenegaard A, Arendt JN, Street A, Gyrd-Hansen D. 2011. Exploring the relationship between costs and quality: does the joint evaluation of costs and quality alter the ranking of Danish hospital departments? *European Journal of Health Economics* **12**: 541–551.
Hvenegaard A, Street A, Sørensen TH, Gyrd-Hansen D. 2009. Comparing hospital costs: what is gained by accounting for more than a case-mix index? *Social Science & Medicine* **69**(4): 640–647.
Iversen T, Aas E, Rosenqvist G, Häkkinen U. 2015. Comparative analysis of costs in EuroHOPE. *Health Economics* **24**(Suppl. 2): 5–22.
Kalseth B, Anthun KS, Hope Ø, Kittelsen SAC, Persson B. 2011. Spesialisthelsetjenesten i norden. Sykehusstruktur, styringsstruktur og lokal arbeidsorganisering som mulig forklaring på kostnadsforskjeller mellom landene. *Rapport* SINTEF Health Services Research.
Kittelsen SAC, Anthun KS, Kalseth B, Halsteinli V, Magnussen J. 2009. En komparativ analyse av spesialisthelsetjenesten i finland, sverige, danmark og norge: aktivitet, ressursbruk og produktivitet 2005–2007. *Rapport* SINTEF Health Services Research.
Kittelsen SAC, Magnussen J, Anthun KS, Häkkinen U, Linna M, Medin E, Olsen K, Rehnberg C. 2008. Hospital productivity and the Norwegian ownership reform—a Nordic comparative study. *STAKES discussion paper* STAKES.
Kittelsen SAC, Persson BA, Anthun KS, Goude F, Hope Ø, Häkkinen U, Kalseth B, Kilsmark J, Medin E, Rehnberg C, Rättö H. 2015. Decomposing the productivity differences between hospitals in the Nordic countries. *Journal of Productivity Analysis* **43**(3): 281–293.
Kruse M, Christensen J. 2013. Is quality costly? Patient and hospital cost drivers in vascular surgery. *Health Economics Review* **3**(22).
Leng GC, Walsh D, Fowkes FG, Swainson CP. 1999. Is the emergency readmission rate a valid outcome indicator? *Quality in Health Care* **8**: 234–238.
Linna M, Hakkinen U, Peltola M, Magnussen J, Anthun KS, Kittelsen S, Roed A, Olsen K, Medin E, Rehnberg C. 2010. Measuring cost efficiency in the Nordic hospitals—a cross-sectional comparison of public hospitals in 2002. *Health Care Management Science* **13**(4): 346–357.
McKay NL, Deily ME. 2008. Cost inefficiency and hospital health outcomes. *Health Economics* **17**: 833–848.
Medin E, Anthun KS, Hakkinen U, Kittelsen SAC, Linna M, Magnussen J, Olsen K, Rehnberg C. 2011. Cost efficiency of university hospitals in the Nordic countries: a cross-country analysis. *European Journal of Health Economics* **12**(6): 509–519.
Medin E, Goude F, Melberg H, Tediosi F, Belicza E, Peltola M. 2015. European regional differences in all-cause mortality and length of stay for hip fracture patients. *Health Economics* **24**(Suppl. 2): 53–64.
Medin E, Häkkinen U, Linna M, Anthun KS, Kittelsen SAC, Rehnberg C. 2013. International hospital productivity comparison: experiences from the Nordic countries. *Health Policy* **112**(1): 80–87.
Moger TA, Peltola M. 2014. Risk adjustment of health care performance measures in a multinational register-based study—a pragmatic approach to a complicated topic. *SAGE Open Medicine* **2**.
Morey RC, Fine DJ, Loree SW, Retzlaff-Roberts DL, Tsubakitani S. 1992. The trade-off between hospital cost and quality of care. An exploratory empirical analysis. *Medical Care* **30**: 677–698.
Mukamel DB, Zwanziger J, Tomaszewski KJ. 2001. HMO penetration, competition, and risk-adjusted hospital mortality. *Health Services Research* **36**: 1019–1035.
OECD. 2009. Health care quality indicators project: patient safety indicators report 2009. *OECD Health Working Papers*.
Osius G, Rojek D. 1992. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association* **87**(420): 1145–1152.

DOI: 10.1002/hec

Ozimek A, Miles D. 2011. Stata utilities for geocoding and generating travel time and travel distance information. *Stata Journal* **11**(1): 106–119.

Peltola M, Seppälä TT, Malmivaaraa A, Belicza E, Numerato D, Goude F, Fletcher E, Heijink R. 2015. Individual and regional-level factors contributing to variation in length of stay after cerebral infarction in six European countries. *Health Economics* **24**(Suppl. 2): 38–52.

Simar L, Wilson PW. 1998. Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. *Management Science* **44**: 49–61.

Stargardt T, Schreyögg J, Kondofersky I. 2014. Measuring the relationship between costs and outcomes: the example of acute myocardial infarction in German hospitals. *Health Economics* **23**(6): 653–669.

Stukel TA, Fisher ES, Alter DA, Guttman A, Ko DT, Fung K, Wodchis WP, Baxter NN, Earle CC, Lee DS. 2012. Association of hospital spending intensity with mortality and readmission rates in Ontario hospitals. *Journal of the American Medical Association* **307**(10): 1037–1045.

Varabyova Y, Schreyögg J. 2013. International comparisons of the technical efficiency of the hospital sector: panel data analysis of OECD countries using parametric and non-parametric approaches. *Health Policy* **112**(1-2): 70–79.

DOI: 10.1002/hec

Paper IV

Int J Health Econ Manag (2017) 17:83–101
DOI 10.1007/s10754-016-9201-9

RESEARCH ARTICLE

Economic incentives and diagnostic coding in a public health care system

Kjartan Sarheim Anthun[1,2] · **Johan Håkon Bjørngaard**[1,3] · **Jon Magnussen**[1]

Published online: 14 October 2016

Abstract We analysed the association between economic incentives and diagnostic coding practice in the Norwegian public health care system. Data included 3,180,578 hospital discharges in Norway covering the period 1999–2008. For reimbursement purposes, all discharges are grouped in diagnosis-related groups (DRGs). We examined pairs of DRGs where the addition of one or more specific diagnoses places the patient in a complicated rather than an uncomplicated group, yielding higher reimbursement. The economic incentive was measured as the potential gain in income by coding a patient as complicated, and we analysed the association between this gain and the share of complicated discharges within the DRG pairs. Using multilevel linear regression modelling, we estimated both differences between hospitals for each DRG pair and changes within hospitals for each DRG pair over time. Over the whole period, a one-DRG-point difference in price was associated with an increased share of complicated discharges of 14.2 (95 % confidence interval [CI] 11.2–17.2) percentage points. However, a one-DRG-point change in prices between years was only associated with a 0.4 (95 % CI −1.1 to 1.8) percentage point change of discharges into the most complicated diagnostic category. Although there was a strong increase in complicated discharges over time, this was not as closely related to price changes as expected.

Keywords Case-mix · DRG · DRG creep · Funding · Hospitals · Financing

JEL Classification I12 · I13 · I18 · G38 · D22 · I10

✉ Kjartan Sarheim Anthun
Kjartan.Anthun@ntnu.no

[1] Department of Public Health and General Practice, NTNU, Norwegian University of Science and Technology, 7491 Trondheim, Norway

[2] Department of Health Research, SINTEF Technology and Society, Trondheim, Norway

[3] Forensic Department and Research Centre Brøset, St. Olav's University Hospital Trondheim, Trondheim, Norway

Introduction

A number of countries have introduced activity-based payment systems for hospital care by linking all or part of the hospital budget to the number of discharged patients while at the same time adjusting for treatment intensity or patient complexity (case mix). The diagnosis-related group (DRG) is one of the most common systems used to account for case mix. DRGs are widely used for both monitoring and payment purposes. The size of the reimbursement differs between patients, reflecting differences in complexity and thus treatment costs. Patients are categorized in different groups based on diagnosis and procedural codes routinely registered in medical records. For some groups, the DRG system makes the distinction between a "complicated" and an "uncomplicated" patient. While the main diagnosis will be the same, complicated patients will have one or more additional "complicating" secondary diagnoses. Within the resulting pair of DRGs, the complicated group will thus have higher predicted costs and a higher reimbursement. Because personnel in hospitals register information about diagnosis, there is the possibility that a patient is consciously coded to a "complicated" DRG. This is often referred to as "upcoding" or "DRG creep", first defined as "a deliberate and systematic shift in a hospital's reported case mix in order to improve reimbursement" (Simborg 1981). It has also been argued that the introduction of activity-based payment systems will increase the importance of accuracy and completeness in coding (Fisher et al. 1992; O'Reilly et al. 2012). The latter view is shared by the Norwegian government body responsible for the Norwegian DRG system, which defines DRG creep as "patients being coded as more complete, resulting in an increase in case mix index" (translated by the authors from Helsedirektoratet (2011)). Indeed, evidence from the US Medicare system indicated that the introduction of a prospective payment system in 1983 was followed by an increase in the average case mix (Carter and Ginsburg 1985; Ellis and McGuire 1986; Carter et al. 1990; Stern and Epstein 1985; Rosenberg 2001).

In the past decade, there has been a renewed interest in issues related to DRG creep and upcoding. Examining a policy reform in the financing of US Medicare discharges, (Dafny 2005) found a positive association between price differences between complicated and uncomplicated DRGs and the share of discharges in complicated groups. More recently, Barros and Braun (2016) found a positive association between price incentives and upcoding in Portugal.

Responses to price incentives vary between different types of hospitals. In Sweden, the increase in the number of secondary diagnoses registered was larger in hospitals with prospective payment systems than hospitals without prospective payment systems (Serdén et al. 2003). Two studies in the USA found that for-profit hospitals were more likely than non-profit or government-owned hospitals to upcode (Dafny and Dranove 2009; Silverman and Skinner 2004), and also that hospitals in "economic distress" were more likely to upcode (Silverman and Skinner 2004). However, no difference in upcoding between public and private hospitals was found in Italy (Berta et al. 2010).

In a cross-country comparative study, Steinbusch et al. suggest that health systems combining for-profit hospitals with the use of secondary diagnosis criteria for classification, such as in the USA, were more susceptible to upcoding (Steinbusch et al. 2007). In a systematic review, Palmer et al. argued that the effects seen in other countries are similar to those observed in the US system (Palmer et al. 2014). In a theoretical work, Kuhn and Siciliani suggested that the level of auditing of the financing system will influence the perceived risk related to upcoding, and this can also explain differences in levels of upcoding across health systems (Kuhn and Siciliani 2008).

The purpose of this paper is to add to the relatively small literature on upcoding in systems dominated by public hospitals by providing an analysis of coding behaviour in Norway over a period of 10 years. The Norwegian health care system is tax funded, with universal access to services that are largely free at the point of use. Hospitals are predominantly publicly owned and financed through a combination of global budgets and activity-based funding. Activity-based financing was introduced in 1997 utilizing a Nordic version of the DRG system. In the period covered by this study (1999–2008), the share of activity-based funding fluctuated between 40 and 60 %.[1] The period also encompasses a major ownership reform in 2002, where hospital ownership was transferred from 19 county councils to the state (Magnussen et al. 2007).

Analysing coding behaviour in the Norwegian health care sector allowed us to address three questions. First, in a public health care system, the additional income generated from upcoding remains in the hospital. Thus, it will be used to increase the level of activity beyond what was planned, to increase slack (inefficiencies), or it will be saved to finance future investments. It remains uncertain to what extent actors in this public setting will seek to increase income by upcoding. Second, the substantial changes in the degree of activity-based funding during the period studied allowed us to analyse to what extent public hospitals adjust their coding behaviour *in response to changes* in financial incentives. Third, using observations over a period of 10 years allowed us to study any underlying trends in coding behaviour, and isolate this from the effects of changes in financial incentives. In all three questions, our main interest was the potential relationship between economic incentives and coding behaviour on an aggregate national level. Although there are numerous micro-level examples of upcoding (Lægreid and Neby 2012; Neby et al. 2015), it is unclear whether these are exceptions to the rule, or whether they represent a general behavioural response to economic incentives.

Materials and methods

Data material

Data from all Norwegian somatic hospital discharges for the period 1999–2008 were used. The Norwegian Patient Registry provided the data.[2] Each hospital discharge was grouped in a DRG, and 250 of the total of 913 groups were linked in complicated/uncomplicated pairs (in 2008). Only patients in acute care hospitals grouped within these 125 DRG pairs were included. We excluded DRG pairs not used in all years, DRG pairs with fewer than 1000 annual cases, and five additional DRG pairs that were viewed as problematic.[3] After exclusion criteria were applied, 3,180,578 in-patient discharges remained. They were grouped into 76 different DRG pairs, of which 53 pairs were medical DRGs and 23 pairs were surgical DRGs.

[1] In 1999–2001, the share of income related to activity was 50 %, increasing to 55 % in 2002 and 60 % in 2003. The share fell to 40 % in 2004, and rose again to 60 % in 2005. The share returned to 40 % in the years 2006–2008.

[2] The Norwegian Patient Register is a complete registry of all specialized hospital care. The interpretation and reporting of these data are the sole responsibility of the authors, and no endorsement by the Norwegian Patient Register is intended nor should be inferred.

[3] These five excluded DRGs were 372/373 (Vaginal births), 76/77 (Other respiratory operating room procedures), 452A/453A (Complications of treatment with surgery), 454/455 (Other injury, poisoning & toxic effect) and 478/479 (Other vascular procedures). Among these DRG pairs, vaginal births was the largest of all complicated/uncomplicated pairs, and was excluded due to significant alterations in the specifications of the DRG pair during the period.

These pairs amount to about 29 % of the total volume of discharges. See Table 1 for a list of included DRG pairs. Our study included 26 hospitals (including three large publicly funded non-profit private hospitals). Not all hospitals treated patients in all included DRGs.

Dependent variable

The dependent variable (c_{tih}) was the percentage of complicated discharges in a DRG pair. This was defined as the number of complicated cases divided by the total number of cases in the DRG pair, calculated for year t, DRG pair i and hospital h.

Potential gain in income from upcoding: the incentive

We measured the potential gain in income from upcoding as the difference in reimbursement (DRG prices) between complicated and uncomplicated groups in each DRG pair similarly to the *spread* in weights as defined by Dafny (2005) and Barros et.al. (Barros and Braun 2016). This spread did not differ across hospitals, as there were no hospital-specific prices. We calculated the difference between prices of complicated and uncomplicated groups within a DRG pair across the years, multiplied by the share of activity-based funding for each specific year. However, we depart from Dafny's approach by calculating the mean across years for each DRG pair and denote this as $\overline{p_i}$ (Eq. 1). To enable comparison across years, we measured prices normalized in DRG points, not as the monetary value of a DRG point. One DRG point, roughly equalling the treatment cost of the "average patient", was valued at 33,647 NOK (∼3629 EUR) in 2008. This should be interpreted as the incentive in a DRG pair because it increases income without increasing cost, should any upcoding take place.

$$\overline{p_i} = \frac{1}{10} * \sum_{t=1999}^{2008} \Big(COMPLICATED_{it} - UNCOMPLICATED_{it} \Big) * ABFSHARE_t \quad (1)$$

In Eq. 1, $COMPLICATED_{it}$ is the DRG weight (relative price) of the complicated group in DRG pair i in year t, $UNCOMPLICATED_{it}$ is the DRG weight of the uncomplicated group in DRG pair i in year t and $ABFSHARE_t$ is the share of the total budget allocated through activity-based financing (from 0 to 1) in year t.

However, the price of each DRG may change from year to year. Such changes are caused by (1) changes in relative reimbursement rates (prices are adjusted annually) for specific DRGs (i.e., $COMPLICATED_{it}$ and $UNCOMPLICATED_{it}$), and (2) variations in the share of activity-based funding between years ($ABFSHARE_t$). Either of these causes will yield changes in the potential gain in income. In this study, we are not only interested in the level of the incentive, ($\overline{p_i}$), but also in changes calculated as the annual changes from the average for each DRG pair (Eq. 2).

$$\Delta p_{it} = \Big(\Big(COMPLICATED_{it} - UNCOMPLICATED_{it} \Big) * ABFSHARE_t \Big) - \overline{p_i} \quad (2)$$

By separating $\overline{p_i}$ and Δp_{it}, we separate the effect of the *level* of the incentive from *changes* in the incentive on coding behaviour. The level of the incentive is thus the difference *between* DRG pairs ($\overline{p_i}$), while the changes are differences over time *within* a specific DRG pair (Δp_{it}). The spread used by Dafny (2005) and Barros et.al. (Barros and Braun 2016) is the sum of these between and within effects.

Table 1 List of DRGs included in study

DRG code	DRG text	M/S	% compl.	# disch. (1000)	Case-mix adjusted # disch. (1000)	$\overline{p_i}$	Mean absolute Δp_{it}
10	Nervous system neoplasm	M	43.1	33.7	42.8	0.379	0.068
18	Cranial and periferal nerve disorders	M	27.5	22.3	17.5	0.230	0.042
24	Seizure and headache age >17	M	25.9	77.9	42.0	0.223	0.075
31	Concussion, age >17	M	16.5	44.4	12.9	0.051	0.022
34	Other disorders of nervous system	M	23.8	78.4	62.3	0.257	0.069
46	Other disorders of the eye, age >17	M	25.3	21.1	10.1	0.256	0.051
68	Otitis media and uri, age >17	M	25.6	24.7	11.7	0.138	0.024
70	Otitis media and uri, age 0–17	M	14.7	34.3	12.3	0.143	0.057
79	Respiratory infections and inflammations, age >17	M	67.7	29.8	61.1	0.390	0.049
89	Simple pneumonia and pleurisy, age >17	M	71.4	186.5	264.5	0.310	0.037
91	Simple pneumonia and pleurisy, age 0–17	M	23.2	18.1	14.4	0.343	0.069
96	Bronchitis and asthma, age >17	M	37.8	25.7	20.3	0.184	0.030
98	Bronchitis and asthma, age 0–17	M	10.1	48.7	28.8	0.204	0.041
99	Respiratory signs and symptoms	M	25.9	26.0	10.9	0.172	0.042
101	Other respiratory system diagnoses	M	40.1	13.3	9.6	0.220	0.029
110	Major cardiovascular procedures	S	55.7	18.2	82.1	0.467	0.179
124	Diagnostic percutan cardiac procedure w circulatory complex dx	M	31.8	33.7	19.0	0.187	0.044
130	Peripheral vascular disorders	M	46.1	58.0	49.1	0.194	0.036
132	Atherosclerosis	M	57.9	43.8	26.9	0.144	0.013
135	Cardiac congenital and valvular disorders age >17	M	73.0	19.1	16.4	0.208	0.066
138	Cardiac arrythmia and conduction disorders	M	35.5	123.9	56.7	0.170	0.033

Table 1 continued

DRG code	DRG text	M/S	% compl.	# disch. (1000)	Case-mix adjusted # disch. (1000)	$\overline{p_i}$	Mean absolute Δp_{it}
141	Syncope and collapse	M	35.5	49.8	21.5	0.078	0.013
144	Other circulatory system diagnoses	M	53.7	23.1	21.4	0.243	0.056
146	Rectal resection	S	54.8	11.9	43.8	0.552	0.149
148	Major small and large bowel procedures	S	59.8	46.6	173.8	0.769	0.158
157	Minor intestinal procedure	S	17.0	30.8	20.0	0.361	0.050
159	Hernia procedures except inguinal and femoral, age >17	S	25.4	12.5	11.3	0.361	0.086
161	Inguinal and femoral hernia procedures, age >17	S	26.0	22.7	14.9	0.154	0.067
170	Other digestive system o. r. procedures	S	40.8	14.2	30.0	0.711	0.170
172	Digestive malignancy	M	68.4	78.6	88.0	0.204	0.047
174	G. i. hemorrhage	M	57.5	51.6	43.1	0.202	0.029
177	Uncomplicated peptic ulcer	M	44.0	10.3	7.6	0.212	0.076
180	G. i. obstruction	M	41.4	15.3	8.5	0.182	0.037
182	Esophagitis, gastroent and misc digest disorders, age >17	M	30.4	249.1	116.0	0.137	0.020
184	Esophagitis, gastroent and misc digest disorders, age 0–17	M	15.8	71.0	26.2	0.103	0.028
188	Other digestive system diagnoses, age >17	M	36.4	41.0	22.7	0.237	0.024
205	Disorders of liver except malig, cirr, alc hepa	M	41.3	17.9	17.8	0.367	0.110
207	Disorders of biliary tract	M	35.1	49.1	36.5	0.243	0.043
210	Hip and femur procedures except major joint, age >17	S	54.9	92.5	189.5	0.302	0.092
218	Lower extrem and humer proc except hip, foot, femur age >17, with cc	S	19.5	55.9	77.4	0.668	0.119
221	Knee procedures	S	13.6	35.8	38.6	0.696	0.172

Table 1 continued

DRG code	DRG text	M/S	% compl.	# disch. (1000)	Case-mix adjusted # disch. (1000)	$\overline{p_i}$	Mean absolute Δp_{it}
223	Major shoulder/elbow proc, or other upper extremity proc	S	13.8	56.2	49.8	0.283	0.048
226	Soft tissue procedures	S	12.4	29.5	21.9	0.421	0.042
228	Major thumb or joint proc, or oth hand or wrist proc	S	22.8	29.1	18.0	0.192	0.087
244	Bone diseases and specific arthropathies	M	37.2	22.1	15.8	0.179	0.028
250	Fracture, sprain, strain or dislocation of forearm, hand or foot, age >17	M	24.2	14.9	5.1	0.214	0.040
253	Fracture, sprain, strain or dislocation of upper arm or lower leg excluding foot, age>17	M	25.5	41.9	22.4	0.234	0.035
257	Total mastectomy for malignancy	S	33.2	15.1	18.2	0.110	0.026
259	Subtotal mastectomy for malignancy	S	22.1	16.2	13.9	0.116	0.011
269	Other skin and subcut tiss proc	S	34.3	21.6	21.2	0.610	0.055
272	Major skin disorders	M	54.5	17.7	24.0	0.307	0.127
277	Cellulitis age >17	M	39.0	45.6	41.7	0.217	0.016
280	Trauma to the skin and subcut tiss age >17	M	34.4	39.9	16.3	0.153	0.021
283	Minor skin disorders	M	25.7	24.2	17.8	0.246	0.074
296	Nutritional and misc metabolic disorders, age >17	M	53.5	27.8	21.9	0.193	0.027
300	Endocrine disorders	M	38.3	20.7	15.6	0.241	0.035
308	Minor bladder procedures	S	26.9	18.9	24.3	0.395	0.278
310	Transurethral procedures	S	37.1	36.3	29.8	0.170	0.040
318	Kidney and urinary tract neoplasms	M	69.6	25.5	31.6	0.365	0.073
320	Kidney and urinary tract infections age >17	M	53.4	71.5	65.8	0.182	0.023
323	Urinary stones, &/or esw lithotripsy	M	29.2	44.9	23.2	0.125	0.031

Table 1 continued

DRG code	DRG text	M/S	% compl.	# disch. (1000)	Case-mix adjusted # disch. (1000)	$\overline{p_i}$	Mean absolute Δp_{it}
325	Kidney and urinary tract signs and symptoms age >17	M	45.8	19.9	9.4	0.108	0.020
331	Other kidney and urinary tract diagnoses age >17	M	47.0	18.2	13.3	0.281	0.070
336	Transurethral prostatectomy	S	40.9	37.4	40.1	0.137	0.020
346	Malignancy, male reprocuctive system	M	72.9	43.5	42.5	0.200	0.056
358	Uterine and adnexa proc for ovarian or adnexal non-malignancy	S	14.4	66.6	90.6	0.429	0.080
366	Malignancy, female reproductive system	M	60.7	47.1	54.2	0.367	0.059
370	Cesarean section	S	31.0	87.5	126.2	0.295	0.069
383	Other antepartum diagnoses w medical complications	M	56.8	56.4	27.0	0.112	0.014
398	Reticuloendothelial and immunity disorders	M	40.5	14.7	14.3	0.320	0.066
403	Lymphoma and non-acute leukemia	M	54.2	72.2	96.6	0.529	0.054
442	Other o. r. procedures for injuries	S	52.2	10.4	28.7	1.192	0.294
444	Traumatic injury, age >17	M	34.9	10.8	5.5	0.241	0.033
449	Poisoning and toxic effects of drugs, age >17	M	29.2	55.2	18.7	0.155	0.043
463	Signs and symptoms	M	36.6	16.0	11.6	0.179	0.042
493	Laparoscopic cholecystectomy w/o c. d. e.	S	25.3	43.8	80.0	0.262	0.043

DRG code and DRG text is for complicated group in the pair
M/S: M = Medical DRG pair, S = Surgical DRG pair
% compl: Percentage of complicated discharges in pair. Defined as number of complicated discharges divided by total number of discharges
# disch: Number of inpatient discharges in DRG pair, 1000
Case-mix adjusted # disch: Case-mix adjusted number of inpatient discharges in DRG pair, 1000 (adjusted by the weights used for reimbursements)
$\overline{p_i}$: Mean difference in prices of complicated and uncomplicated group in pair
Mean absolute Δp_{it}: Mean absolute deviation from $\overline{p_i}$. Since the mean deviation from the mean in a group always is zero, we have here showed the mean absolute deviation in this table

Statistical analysis

The clustered and hierarchical nature of the data led us towards a mixed-model approach. The multivariable analyses were performed using a three-level linear regression model, where hospital discharges were aggregated to 19,250 observations, comprising 10 yearly observations (level 1) of each DRG pair (level 2) within each of the 26 hospitals (level 3). Equation 3 describes our main analytical model.

$$c_{tih} = a + a_i + a_h + b_1\overline{p_i} + b_2\Delta p_{it} + b_3 T_t + b_4 D + b_5 T_t D + b_x x_{tih} + \varepsilon_{tih} \qquad (3)$$

Our dependent variable, c_{tih}, is the share of complicated cases in year t in DRG pair i in hospital h. The effects of the level of the upcoding incentive were defined by $\overline{p_i}$ (Eq. 1), and the change in incentive defined by Δp_{it} (Eq. 2). To capture any general development in coding practice over time, we included time trend (T_t), which measures years since 1999. This time trend might, however, capture both general improvements in quality of coding, as well as any fraudulent upcoding not captured by the effects of $\overline{p_i}$ and Δp_{it}. We also controlled (by way of a dummy (D) for the years 2002–2008) for the possible effect of the ownership reform in 2002. A statistical interaction of these was included ($T_t D$).

The a-terms are constants and intercepts at the different levels while ε_{tih} is the residual. Other covariates are denoted x_{tih} in the equation. These included average age and sex in each DRG pair. Elderly patients are more likely to be frailer, and therefore have an increased probability of being grouped in complicated groups.[4] For the same reason, we also adjusted for emergency status and length of stay. Emergency admissions are more likely to be complicated than elective procedures (Melnick et al. 1989; Keller et al. 1987). Length of stay may be a proxy for case mix as the longer the patient remains in the hospital, the more complex the illness is likely to be or the frailer the patient. To better control for co-morbidity and case mix, we constructed a Charlson index for each analytical observation. The index is a measure of co-morbidity that is based upon secondary diagnoses (Charlson et al. 1987), as also was our dependent variable. For the calculation of the Charlson index, we excluded those diagnoses that caused a complicated DRG grouping (within each DRG pair), and thus the index does not have an upcoding bias other than what comes from the complicated discharges actually being more complicated.

While ownership of hospitals after 2002 was transferred to the state, there was an administrative decentralization to four regional health authorities. The regional health authorities face different challenges, as there are substantial differences in distance to hospital, different degrees of deficits/surpluses and also size of population. We also included dummy variables for these to account for possible regional variances in coding behaviour induced by diverse organizational incentives or structures. The annual number of in-patient treatments at each hospital (measured as case mix-adjusted DRG points) was included as a proxy for hospital size. This measure will be invariant at the DRG pair level. Finally, we performed a stratified analysis of medical and surgical DRGs, because surgical DRGs could arguably have less room for differences in coding behaviour than medical DRGs. Precision was estimated with 95 % confidence intervals (CI).

Even though the dependent variable is a proportion, we assumed normality in the residuals. Robustness tests were performed with a simpler two-level model, using the actual monetary value as main independent variables instead of the rather abstract DRG points.

[4] In the regressions, we control for age by restricted cubic splines, calculated with five knots (Harrell 2001). Five knots means that the age range is split in five groups. These splines provide a better control and fit of variables than a simple linear approach. However, the resulting coefficients are not readily interpretable as they are not marginal linear effects.

Table 2 Descriptive statistics for variables in analysis

Variable	Mean	Median	Std. Dev.	Min	Max
Age	55.57	58.16	1.59	1.00	98.00
Percentage female	51.19	49.70	21.09	0.00	100.00
Percentage emergency	70.75	81.24	29.16	0.00	100.00
Length of stay	4.87	4.10	3.10	0.00	46.00
Number of inpatient treatments at hospital*	11,496	8959	8383	1812	43,540
Percentage medical DRGs	70.20	100.00	45.73	0.00	100.00
Charlson co-morbidity index	0.26	0.18	0.33	0.00	8.00
Potential gain in income $\overline{p_i}$	0.28	0.23	0.18	0.05	1.19
Changes in potential gain in income Δp_{it}	0.00	−0.00	0.09	−0.33	0.52
Percentage complicated discharges (c_{tih})	38.01	35.30	20.94	0.00	100.00

$N = 19{,}250$
*Case-mix adjusted, DRG-pair invariant

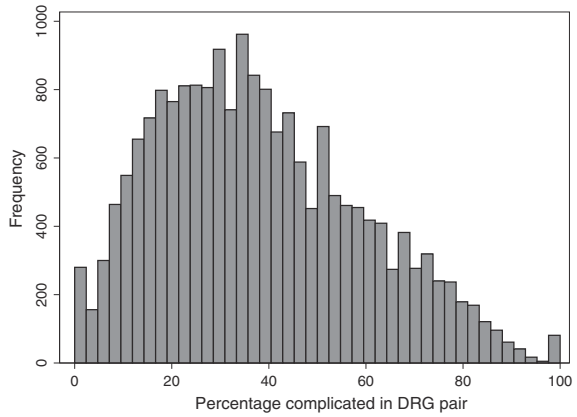


Fig. 1 Distribution of percentage complicated in DRG pair, histogram

Results

Descriptive statistics

Table 2 presents descriptive statistics. Across the observations (year, DRG pair, hospital), the mean share of complicated discharges was 38 %, ranging from 0 to 100 (see Fig. 1 for distribution). The mean $\overline{p_i}$ was 0.28 DRG points and ranged from 0.05 to 1.19 (see Fig. 2 for distribution). The mean change (Δp_{it}) was zero because this was defined as yearly deviations from $\overline{p_i}$. Table 1 lists $\overline{p_i}$ and the mean absolute Δp_{it} for each DRG pair, and Fig. 3 shows the distribution of Δp_{it}.

Data analysis was performed at an aggregate level, i.e., the mean age of 55.6 was the mean across all observations (year, DRG pair, hospital) and not the mean for all distinct patients.

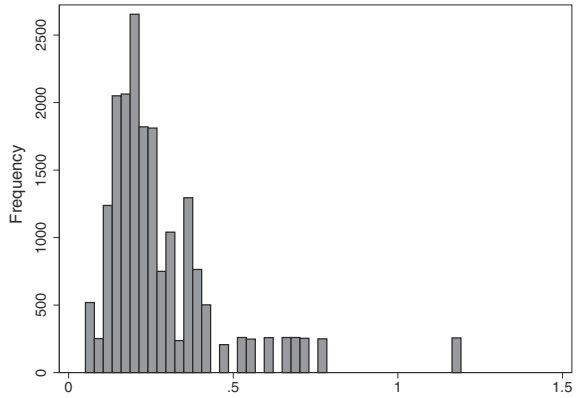


Fig. 2 Distribution of potential gain in income $\overline{p_i}$, histogram

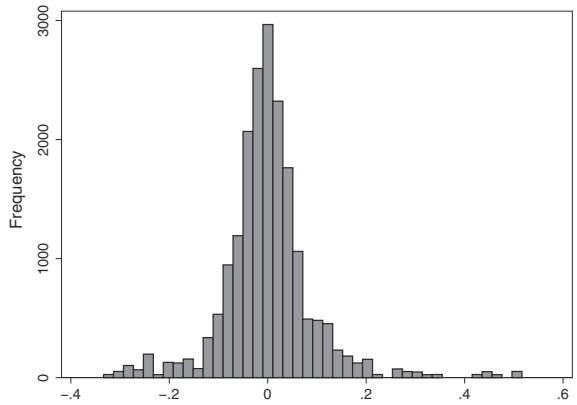


Fig. 3 Distribution of changes in potential gain in income Δp_{it}, histogram

On average, the share of females was 51.2 %, but this varied from 0 to 100 as some DRG pairs were gender specific. The mean length of stay was 4.87, but varied across DRG pairs with a maximum of 46. Some DRG pairs had a zero length of stay and were thus likely to be patients admitted as in-patients but discharged on the same day. There was a downward trend in length of stay over the period. To control for hospital size, we also calculated the (case mix-adjusted) number of in-patient discharges at each hospital. This was measured annually at the hospital level, and as opposed to the other independent variables, this was DRG pair invariant. Hospital size varied substantially with the mean of 11,496 discharges while the largest hospital had 43,540 discharges. Mean hospital size also increased over the period covered by this study, both through reforms and reorganizations/mergers as well as increased budgets. All control variables were centred on their mean in the multivariable analysis.

Multivariable analysis

Table 3 shows the correlations between the variables of interest. The share of complicated discharges (c_{tih}) was highly correlated with the case mix-related variables: age (Pearson's r correlation coefficient 0.512), length of stay (0.461) and comorbidity (0.510). The share of complicated discharges was also positively correlated with the temporal variables, emergency admissions and medical DRG pairs. At this aggregate level, there was a small yet statistically significant association with $\overline{p_i}$ (0.091), but not with Δp_{it}.

In the multilevel regressions, there was a positive association between $\overline{p_i}$ and the share of complicated discharges (Table 4). Over the whole period, a one-DRG-point difference in $\overline{p_i}$ was associated with an increased share of complicated discharges of 14.2 percentage points (95 % CI 11.2–17.2). However, a one-DRG-point change in Δp_{it} between years was only associated with an increase of the most complicated group of 0.4 percentage points (95 % CI −1.1 to 1.8).

The temporal variables had large estimated values. There was a large annual increase in the share of complicated discharges of 2.9 percentage points (95 % CI 2.6–3.1) in the period leading up to the reform (1999–2001). After the reform in 2002, there was a shift in the share of complicated discharges of 10.2 percentage points (95 % CI 9.6–10.8). By calculating the combined estimates of T_t, D and $T_t D$, we find an annual increase of only 0.4 percentage points in the period after 2002.

The case-mix adjustors had a large impact on the share of complicated discharges. A one-unit increase in the Charlson index, which can be interpreted as one more co-morbidity, was associated with an increase of 12.5 percentage points in the share of complicated discharges. For an increase in mean length of stay of one day, the share of complicated discharges increased 1.3 percentage points (95 % CI 1.2–1.4). We found only a small negative association between share of females and percentage of complicated discharges. There were no substantial differences between the different regional health authorities. Hospital size had a small positive effect, indicating that larger hospitals have a higher share of complicated discharges.

The share of complicated discharges was 8.1 percentage points (95 % CI 6.8–9.4) higher in medical DRG pairs than in surgical DRG pairs. We performed a stratified analysis of medical and surgical DRG pairs. For medical pairs, a one-DRG-point change in Δp_{it} was associated with an increase in share of complicated discharges of 5.1 percentage points (95 % CI 2.5–7.6) (Table 4); for the surgical DRG pairs, there was a negative effect from Δp_{it} of −2.5 (95 % CI −4.3 to −0.6). Aside from the effect of Δp_{it}, there were no other large differences between the stratified and the non-stratified analyses.

Robustness tests were performed using simpler two-level models (either hospital level or DRG pair level), but the results did not differ much from the results presented in Table 4. We also ran the analysis using potential income gain measures calculated from the monetary refund that the hospitals received instead of DRG points. The refund was calculated using the yearly refund value of a DRG point while deflating the older years to real 2008 prices. The results did not differ much from the presented results. The test showed that for every 1000 NOK (∼109 EUR) in increased potential income ($\overline{p_i}$), the share of complicated discharges increased by 0.31 percentage points. Nonetheless, changes in Δp_{it} had no effect. Table 5 shows the different models tested for robustness.

Table 3 Correlation matrix of share complicated discharges and all independent variables

	Percentage complicated discharges c_{tih}	$\overline{p_i}$	Δp_{it}	Time trend (T_t)	Reform (shift 2002–2008)	Interaction time trend and reform	Age	Share female patients	Share emergency admissions	Length of stay	Hospital size	Medical DRG pairs (dummy)
$\overline{p_i}$	0.091∗											
Δp_{it}	0.005	0.000										
Time trend (T_t)	0.246∗	−0.000	−0.058∗									
Reform (shift 2002–2008)	0.258∗	−0.000	0.011	0.798∗								
Interaction time trend and reform	0.249∗	−0.000	−0.0482∗	0.984∗	0.854∗							
Age	0.512∗	−0.003	−0.001	0.022∗	0.021∗	0.022∗						
Share female patients	−0.064∗	0.007	0.003	−0.003	0.002	−0.003	−0.099∗					
Share emergency admissions	0.163∗	−0.215∗	−0.005	0.048∗	0.030∗	0.047∗	−0.116∗	−0.059				
Length of stay	0.461∗	0.514∗	0.032∗	−0.116∗	−0.101∗	−0.114∗	0.347∗	0.051∗	−0.146∗			
Hospital size	0.015∗	0.001	−0.001	0.070∗	0.072∗	0.070∗	−0.099∗	−0.017∗	−0.118∗	0.010		
Medical DRG pairs (dummy)	0.197∗	−0.497∗	−0.000	0.006	0.004	0.006	−0.035∗	−0.091∗	0.650∗	−0.224∗	−0.013	
Mean Charlson index	0.510∗	0.0737∗	0.006	0.154∗	0.140∗	0.153∗	0.355∗	−0.053∗	0.018∗	0.354∗	0.026∗	0.100*

$* p < 0.05$

Table 4 Multilevel linear regression of the percentage of complicated discharges, coefficients with 95 % CI in parenthesis

	Complete model	Only surgical DRG pairs	Only medical DRG pairs
Potential gain in income $\overline{p_i}$	14.23*** (11.23 to 17.24)	17.08*** (14.21 to 19.95)	13.19*** (6.09 to 20.29)
Changes in potential gain in income Δp_{it}	0.35 (−1.10 to 1.79)	−2.45*** (−4.27 to −0.62)	5.08*** (2.54 to 7.63)
Time trend (years since 1999)	2.85*** (2.58 to 3.12)	3.04*** (2.52 to 3.57)	2.85*** (2.54 to 3.17)
Reform (dummy for years 2002–2008)	10.23*** (9.64 to 10.81)	9.66*** (8.55 to 10.77)	10.60*** (9.92 to 11.27)
Interaction time trend and reform	−2.41*** (−2.68 to −2.13)	−2.49*** (−3.02 to −1.96)	−2.42*** (−2.74 to −2.10)
Ten percentage points increase in women	−0.22*** (−0.35 to −0.09)	−0.41*** (−0.60 to −0.21)	−0.20** (−0.36 to −0.04)
Ten percentage points increase in emergency admissions	0.96*** (0.83 to 1.09)	0.62*** (0.42 to 0.81)	1.15*** (0.96 to 1.33)
Length of stay	1.25*** (1.16 to 1.35)	1.29*** (1.14 to 1.43)	1.25*** (1.12 to 1.38)
Hospital size (case-mix adjusted number of inpatient treatments/1000)	0.55*** (0.42 to 0.68)	0.38*** (0.22 to 0.53)	0.42*** (0.29 to 0.55)
Medical DRG pairs compared with surgical	8.09*** (6.78 to 9.40)		
Charlson index	12.54*** (11.74 to 13.34)	10.44*** (8.91 to 11.97)	13.59*** (12.66 to 14.53)
N	19,250	5,736	13,514

*** $p < 0.01$,** $p < 0.05$,* $p < 0.1$. Controlled for regional health authorities (with dummies) and five age splines. Random effects of time trend, otherwise fixed effects

Table 5 Multilevel linear regression of the percentage of complicated discharges; robustness test of other specifications: two level model combined DRG pairs and hospital, two level model DRG pairs, and monetary value of price incentive and changes in price incentive, $N = 19,250$

	Two level model hospital	Two level model DRG pairs	Monetary value
Potential gain in income $\overline{p_i}$	10.67*** (9.27 to 12.06)	15.21** (2.99 to 27.43)	0.31*** (0.24 to 0.38)
Changes in potential gain in income Δp_{it}	−0.09 (−1.98 to 1.81)	0.73 (−1.23 to 2.68)	0.02 (−0.02 to 0.05)
Time trend (years since 1999)	3.11*** (2.62 to 3.59)	2.88*** (2.51 to 3.26)	2.98*** (2.71 to 3.26)
Reform (dummy for years 2002–2008)	10.64*** (9.78 to 11.50)	10.56*** (9.87 to 11.26)	10.48*** (9.89 to 11.06)
Interaction time trend and reform	−2.51*** (−2.92 to −2.10)	−2.48*** (−2.81 to −2.14)	−2.45*** (−2.72 to −2.17)
Ten percentage points increase in women	0.28*** (0.19 to 0.37)	−0.36*** (−0.51 to −0.20)	−0.22*** (−0.35 to −0.09)
Ten percentage points increase in emergency admissions	0.71*** (0.62 to 0.79)	0.85*** (0.75 to 0.96)	0.96*** (0.83 to 1.10)
Length of stay	1.99*** (1.91 to 2.08)	1.16*** (1.07 to 1.25)	1.26*** (1.17 to 1.36)
Hospital size (case-mix adjusted number of inpatient treatments/1000)	0.29*** (0.14 to 0.44)	0.09*** (0.07 to 0.11)	0.55*** (0.43 to 0.68)
Medical DRG pairs compared with surgical	10.44*** (9.83 to 11.05)	7.22*** (2.32 to 12.13)	7.91*** (6.61 to 9.20)
Charlson index	9.77*** (9.13 to 10.41)	16.27*** (15.42 to 17.11)	12.57*** (11.77 to 13.37)

*** $p < 0.01$,** $p < 0.05$,* $p < 0.1$. Controlled for regional health authorities (with dummies) and five age splines. Random effects of time trend, otherwise fixed effects

Discussion

Our goal was to examine the association between the potential gain in income from upcoding and the coding behaviour of hospitals. Across DRG pairs, we found a positive association between the gain in income from upcoding and the share of discharges classified as complicated. Thus, DRG pairs in which there was a higher gain in income from upcoding also had a higher share of complicated discharges. However, although we controlled for co-morbidity, age and length of stay, we cannot exclude the possibility that this partly reflects differences in the case mix. Nevertheless, it is not clear why the difference in treatment costs between complicated and uncomplicated discharges should be higher in DRG pairs with a higher share of complicated discharges and therefore our results indicate that coding behaviour is related to the size of the incentive.

We found that a difference in price between a complicated and uncomplicated group of one DRG point was related to a difference of 14 percentage points in the share of complicated discharges within a DRG pair. Although this may seem like a large effect, the average potential gain from upcoding was only 0.28 DRG points (see Table 2).

We found no association between changes in Δp_{it} over time and the share of complicated discharges within a DRG pair. Thus, in a period with frequent changes in the share of activity-based funding, hospitals did not seem to respond by changing their coding behaviour. However, when stratifying the analysis by medical and surgical DRGs, we found a small, positive association for medical DRGs. Because surgical patients are generally more homogeneous (within a DRG) than medical patients, there may have been less opportunity for tactical coding of these patients. Although the size of the estimated association was small, this result indicated that there might be subgroups of patients where the relationship between financial incentives and tactical coding is stronger. This corresponds to earlier results on how Norwegian hospitals respond to price changes (Januleviciute et al. 2016). Melberg et al. have recently shown higher growth in DRG groups with a price increase than in groups with a reduction in reimbursement rates (Melberg et al. 2016).

We found that the share of complicated discharges increased during the ten year period covered by the study. This may be due to changes in case mix resulting from demographic changes, changes in technology, changes in the quality and completeness of coding and finally changes in the financing system. Recalling the two different definitions of upcoding and DRG creep presented in the introduction, we cannot here distinguish between "deliberate upcoding" and "more complete coding". The increasing trend could both indicate that the quality of coding has improved, and at the same time that the presence of explicit and implicit incentives is followed by a general increase in the recording of secondary diagnoses. Thus, while we cannot label all upcoding as being completely driven by financial incentives, we argue that such incentives were present and that their consequences are reflected on an aggregate level by the increasing time trend. The introduction of activity-based funding in 1997 was followed by an increased use of secondary diagnoses. Eventually the use of secondary diagnoses will reach a level (or equilibrium) where it might be difficult to justify an additional secondary diagnosis from a medical point of view. Thus, one might suspect that a large part of the potential for increase was exhausted in the period following the hospital reform, explaining the slowing growth in the share of complicated discharges.

This paper decomposed the price incentive into two components, $\overline{p_i}$ and Δp_{it}, to differentiate between the level and changes of the incentive for upcoding. This approach differs from earlier studies but demonstrates that, in Norway, the differences in prices are more important

than changes within groups. Hospitals may appear to respond to prices, but the changes in price are probably too small to have a large-scale impact.

We believe that the major strength of this analysis is the fact that we are able to utilize a complete dataset covering all DRG pairs for all patients at all hospitals. Our analyses include a ten year period in which there have been large and repeated changes in the potential gain in income from upcoding. Thus, any aggregate effects of increased gain in income from upcoding should be detected in this study. By controlling for a time trend and separating within and between effects, we are more reassured that any remaining effects are more related to upcoding rather than to an increase in the quality of coding.

We have employed a system perspective by pooling all DRG pairs, hospitals and years in the same analysis. This could dilute important findings for specific DRG pairs. Silverman and Skinner (2004) found substantial evidence of upcoding for patients with pneumonia. Their results were robust to different model specifications, but sensitive to the included DRGs. Our stratification showed very different results for the medical and surgical DRG pairs. It is safe to assume that even larger differences will be found on examination of separate DRGs. However, our aim was to detect system-level effects and not effects of singular groups or hospitals. One might also question whether the observed changes in the price incentive were large enough to have an effect. While frequent and potentially substantial, the changes in incentives observed in this study were small compared with some of the larger exogenous shocks described by, for example, Dafny (2005). Therefore, it may have been unrealistic to expect significant results from the observed changes. A change of 20 percentage points in the share of activity-based funding is, however, not trivial and it is interesting that these changes only seem to have led to a marginal change in coding practice.

Upcoding can take place in all systems that incentivize documenting of diagnoses. We have limited our study to upcoding in DRG pairs in Norway. These groups amount to less than one-third of the total volume of treatment. Upcoding is possible for all groups, but the paired structure of complicated/uncomplicated lends itself easily to our research strategy of testing directly whether incentives are associated with upcoding. There are several ways "manipulations" can occur in a DRG system (Neby et al. 2015). In this paper, we have focused solely on upcoding and not touched upon other related strategies: gaming, dumping, skimping and skimming. Further studies should attempt to distinguish upcoding from other manipulations empirically. It is impossible using registry data to determine whether the upcoding has been deliberate. To assess the actual conscious decision to upcode, one must opt for a qualitative approach. This study has not ventured into the auditing of diagnosis and hospital records. Earlier evidence from Norway has indicated that diagnostic accuracy is not very high (Jørgenvåg 2005), and it would be interesting to consider whether the Norwegian auditing scheme could be considered optimal (Kuhn and Siciliani 2008).

Acknowledgements The authors wish to thank Hugh Gravelle, Line Planck Kongstad and Søren Rud Kristensen for feedback on earlier drafts of the paper, and also to thank the participants on the Nordic Health Economics' Study Group annual meetings in Oslo 2013 and Reykjavik 2014, as well as the participants on the EuHEA PhD and Early Career Researcher Conference in Manchester 2014. The authors also wish to thank the anonymous reviewers of the journal for helpful comments on earlier version of the paper.

Funding This study was funded by the Norwegian research council (Grant number 214338).

Compliance with ethical standarrd

Conflict of interest The authors declare that they have no conflict of interest.

References

Barros, P., & Braun, G. (2016). Upcoding in a National Health Service: The evidence from Portugal. *Health Economics*. doi:10.1002/hec.3335.

Berta, P., Callea, G., Martini, G., & Vittadini, G. (2010). The effects of upcoding, cream skimming and readmissions on the Italian hospitals efficiency: A population-based investigation. *Economic Modelling*, *27*(4), 812–821.

Carter, G. M., & Ginsburg, P. B. (1985). The medicare case mix index increase: Medical Practice Changes, Aging and DRG Creep. *Rand Publication Series*. Santa Monica: Rand Corporation Report R-3292-HCFA.

Carter, G. M., Newhouse, J. P., & Relles, D. A. (1990). How much change in the case mix index is DRG creep? *Journal of Health Economics*, *9*(4), 411–428.

Charlson, M. E., Pompei, P., Ales, K. L., & MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of chronic diseases*, *40*(5), 373–383.

Dafny, L., & Dranove, D. (2009). Regulatory exploitation and management changes: Upcoding in the hospital industry. *Journal of Law and Economics*, *52*(2), 223–250.

Dafny, L. S. (2005). How do hospitals respond to price changes? *The American Economic Review*, *95*(5), 1525–1547.

Ellis, R. P., & McGuire, T. G. (1986). Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of Health Economics*, *5*(2), 129–151.

Fisher, E. S., Whaley, F. S., Krushat, W. M., Malenka, D. J., Fleming, C., Baron, J. A., et al. (1992). The accuracy of Medicare's hospital claims data: Progress has been made, but problems remain. *American Journal of Public Health*, *82*(2), 243–248.

Harrell, F. E. J. (2001). *Regression modeling strategies: With applications to linear models, logistic regression and survival analysis*. New York: Springer.

Helsedirektoratet (2011). DRG-ordliste [DRG Dictionary]. http://www.helsedirektoratet.no/finansiering/drg/ordliste/. Accessed 01.07.2014.

Januleviciute, J., Askildsen, J. E., Kaarboe, O., Siciliani, L., & Sutton, M. (2016). How do hospitals respond to price changes? Evidence from Norway. *Health Economics*, *25*, 620–636. doi:10.1002/hec.3179.

Jørgenvåg, R. H., Øyvind, B. (2005). Kvalitet på medisinsk koding og ISF-refusjoner. I hvilken grad er journalgjennomgang et nyttig verktøy [Quality of diagnostic coding and activity based financing. To what extent is journal revision a useful tool?]. SINTEF Report STF78 A055501, Trondheim Norway.

Keller, S. M., Markovitz, L. J., Wilder, J. R., & Aufses, A. H. (1987). Emergency and elective surgery in patients over age 70. *The American Surgeon*, *53*(11), 636–640.

Kuhn, M., & Siciliani, L. (2008). Upcoding and optimal auditing in health care (or the economics of DRG creep). *CEPR Discussion Paper No. DP6689*.

Lægreid, P., & Neby, S. (2012). Gaming the system and accountability relations: Negative side-effects of activity-based funding in the Norwegian hospital system. *Stein Rokkan Centre for Social Studies Working Paper: 10-2012*: UNI Rokkan Centre.

Magnussen, J., Hagen, T. P., & Kaarboe, O. M. (2007). Centralized or decentralized? A case study of Norwegian hospital reform. *Social Science & Medicine*, *64*(10), 2129–2137.

Melberg, H. O., Beck Olsen, C., & Pedersen, K. (2016). Did hospitals respond to changes in weights of Diagnosis Related Groups in Norway between 2006 and 2013? *Health Policy*, *120*(9), 992–1000. doi:10.1016/j.healthpol.2016.07.013.

Melnick, G. A., Serrato, C. A., & Mann, J. M. (1989). Prospective payments to hospitals: Should emergency admissions have higher rates? *Health Care Financing Review*, *10*(3), 29–39.

Neby, S., Lægreid, P., Mattei, P., & Feiler, T. (2015). Bending the rules to play the game: Accountability, DRG and waiting list scandals in Norway and Germany. *European Policy Analysis*, *1*(1), 127–148.

O'Reilly, J., Busse, R., Häkkinen, U., Or, Z., Street, A., & Wiley, M. (2012). Paying for hospital care: The experience with implementing activity-based funding in five European countries. *Health Economics, Policy and Law, 7*(Special Issue 01), 73–101. doi:10.1017/S1744133111000314.

Palmer, K. S., Agoritsas, T., Martin, D., Scott, T., Mulla, S. M., Miller, A. P., et al. (2014). Activity-based funding of hospitals and its impact on mortality, readmission, discharge destination, severity of illness,

and volume of care: A systematic review and meta-analysis. *PLOS ONE*, *9*(10), e109975. doi:10.1371/journal.pone.0109975.

Rosenberg, M. A., & Browne, M. J. (2001). The Impact of the inpatient prospective payment system and diagnosis-related groups. *North American Actuarial Journal*, *5*(4), 84–94.

Serdén, L., Lindqvist, R., & Rosén, M. (2003). Have DRG-based prospective payment systems influenced the number of secondary diagnoses in health care administrative data? *Health Policy*, *65*(2), 101–107.

Silverman, E., & Skinner, J. (2004). Medicare upcoding and hospital ownership. *Journal of Health Economics*, *23*(2), 369–389.

Simborg, D. W. (1981). DRG creep: A new hospital-acquired disease. *The New England Journal of Medicine*, *304*(26), 1602.

Steinbusch, P. J., Oostenbrink, J. B., Zuurbier, J. J., & Schaepkens, F. J. (2007). The risk of upcoding in casemix systems: A comparative study. *Health Policy*, *81*(2), 289–299.

Stern, R. S., & Epstein, A. M. (1985). Institutional responses to prospective payment based on diagnosis-related groups: Implications for cost, quality, and access. *Hospital Topics*, *63*(3), 18–24.