INTERNAL MARKETS, MANAGEMENT BY TARGETS, AND QUASI-MARKETS: AN ANALYSIS OF HEALTH CARE REFORMS IN THE ENGLISH NHS

Kristian Niemietz*

Abstract
There have been two major attempts to introduce market mechanisms into England's National Health Service: the 'internal market' reform project of the 1990s, and the 'quasi-market' of the 2000s. Despite their similarities, the former attempt was on balance unsuccessful while the latter succeeded. This article examines and compares the outcomes of the two periods, analysing the reasons for their relative successes and failures. It goes on to highlight options for future reforms that would build on those achievements.

JEL codes: H44, H51, I11, I18.

Keywords: Foundation Trusts; internal market; NHS reform; Payment by Results; purchaser–provider split; quasi-market.

1. Introduction

Supporters of reform of England's National Health Service (NHS) often feel that the institution's status of a 'national religion' makes any reasoned debate about the subject near-impossible. As Taylor (2013, pp. 7–8) puts it:

> We love our health service. We love it in a way that has no parallel in other countries. . . . The social consensus is so strong around the NHS that dissenting voices sound jarring. When [Conservative Member of the European Parliament] Daniel Hannan described the NHS as a 'mistake' on US TV, there was genuine shock and surprise back home. David Cameron described his opinion as 'eccentric'. He was right. People in Britain do not hold views like that.

Taylor's explanation:

> Many people believe that the very idea of universal healthcare . . . is an idea invented in Britain and uniquely realised in Britain. None of this is true. But it leads us to hold the institutions of the NHS in a peculiar reverence.

There is indeed very little serious debate about health care reform in the UK. Political statements on health care amount to little else than an arms race over who professes the

*Senior Research Fellow, Institute of Economic Affairs. Email: kniemietz@iea.org.uk. The author would like to thank the Age Endeavour Fellowship for its generous financial support, which has made possible the research project of which this article is a part. The article is loosely based on a presentation given at the IEA conference on New Frontiers in Privatisation, 31 August–3 September 2014, Lake Bled, Slovenia.

greatest love for the NHS, and who makes the most credible commitment to ring-fence its budget. At general elections voters do not get to choose between competing visions about what kind of health service they want. Rather, it seems that the only position on the NHS that is politically palatable is to shower it with praise, and to promise to spend more money on it. Against this background, one would expect health care in the UK to be a reform-free zone.

But, surprisingly, this is not the case. Since the 1990s there have been two major attempts to introduce market-like mechanisms into the health service: the 'internal market' of the 1990s and the 'quasi-market' of the 2000s. The first attempt was, on balance, unsuccessful, and was largely abandoned in 1997. The second attempt, though, was a qualified success, even if it was clearly 'unfinished business'. This article evaluates both reform periods, and provides suggestions for the next generation of reforms.

2. The internal market of the 1990s

2.1. How the internal market worked

The introduction of market mechanisms into the NHS was first tried in the early 1990s. The idea was to retain a nationalised single funder and monopoly provider, but to simulate market processes within it by creating a degree of internal competition between some of its constituent parts (hence 'internal market'). To this end, the two key functions of the NHS, the allocation of funds and the actual delivery of health care services, were separated; they were henceforth to be performed by distinct subsets of NHS organisations (the 'purchaser–provider split'). District health authorities became internal commissioners. They were given health care budgets from which they were to 'buy' health care services from local NHS providers to meet the health needs of their respective local populations. They were meant to do so selectively, systematically favouring better-performing providers over worse-performing ones, as a purchaser would behave in a real market. Complementary to the establishment of stand-alone purchasing entities, hospitals also became legal entities in their own right ('NHS trusts'), with a degree of autonomy over their own affairs.

District health authorities were initially constituted as local monopolies, but competition was introduced on the purchasing side as well. General practitioners (GPs) were given the ability to partly opt out of their district health authority's commissioning arrangements, and replace them with their own. These 'general practice fundholders' would be assigned their own commissioning budgets, and become partial purchasers of secondary and tertiary health care for the patients registered with their practice.

The internal market lacked most of the defining features of an actual market. For a start, there was no market-determined entry and exit of providers (Propper, Burgess and Gossage 2008, p. 145). Underperforming hospitals were not be allowed to fail, which is why the internal market has been likened to 'the caucus race in *Alice in Wonderland*, in which "everyone must have prizes" ' (Bevan and Hamblin 2009, p. 162). As the flipside of the same coin, well-performing hospitals were constrained in their ability to expand; indeed, hospitals were not even allowed to retain budget surpluses. The distribution of hospital facilities across the country was still determined by national plans, not competition. Competition therefore remained mostly a 'dry run', the outcomes of which would have no serious consequences. Commissioners also took the attribute 'internal' quite literally: the NHS remained a closed shop, which barely

interacted with independent providers (Propper, Burgess and Green 2004, p. 1249). NHS providers never had to fear competitive pressure from newcomers or outsiders. Patient choice was not really part of the internal market experiment either. Patients could choose their GP within a catchment area, and they could indirectly exert choice in secondary and tertiary care by choosing between GPs with different referral patterns. But there was no patient choice at the point of referral.

The internal market of the 1990s is best described as a managerialist version of competition, that is, competition between different parts of an organisation that were all operating under the same central guidelines. Competition was meant to act as a spur, not as a discovery process in a Hayekian sense. Indeed, there was nothing to discover: it was competition between participants who were all doing the same thing. But it was still a break with the cosy competition-free world of the past.

2.2. Where the internal market succeeded

Empirical evidence about the impact of the internal market is quite mixed. GP fundholding, the scheme which gave GPs the option to commission elements of secondary or tertiary health care themselves rather than relying on their district health authority, was probably the most successful element. Propper, Croxson and Shearer (2002) tested whether GP fundholders managed to cut hospital waiting times for their patients by optimising their referral patterns. They looked at the difference-in-differences[1] in the evolution of waiting times, comparing patients whose GPs became fundholders with patients whose GPs did not, while controlling for patient characteristics. They also compared waiting times for treatments which fundholders were allowed to commission themselves, and waiting times for treatments over which fundholders had no control.

The authors found that fundholders did achieve shorter waiting times, but the effect was limited to fundholders' own patients and to those treatments over which fundholders had direct control. There were no spillover effects, that is, no evidence that fundholders managed to speed up hospital access across the board. One important channel through which competition raises standards in other markets was therefore missing.

> Our results indicate that the scheme led to some improvement in the quality of service provided, but only for a limited set of patients and a limited set of treatments. . . . However, because fundholders' patients having non-fundholding procedures did not gain, the overall average waits of fundholders patients were not significantly less than those of non-fundholders. (Propper, Croxson and Shearer 2002, p. 249)

The conversion of health authority-managed hospitals into more autonomous NHS trusts was another element of the internal market which seems to have worked rather well. Söderlund et al. (1997) compared average costs per inpatient episode in different hospitals over time, and controlled for factors such as patient case mix, regional differences in wage and capital costs, and various hospital characteristics. The authors exploited the fact that the conversion occurred in stages, and that some hospitals remained under direct management of the health authorities throughout, acting as a quasi-control group. They found that the change in status was associated with productivity improvements:

> In the longitudinal model . . . trust status . . . had a significant negative effect on average costs. The period dummy variables in both models indicate that overall productivity of the hospitals improved over the three years and that improvements . . . were significant at the 5% level. (Söderlund et al. 1997, p. 1127)

Inadvertently, the study by Söderlund et al. also illustrated how far away the internal market was from a real market. Almost all of the control variables used in this model are variables that, in a proper market, would themselves have been *determined by* competition – they would have been competitive outcomes, not exogenous controls. Competition, properly understood, should be a means to *discover* on what scale a hospital should operate, how specialised it should be, what its specialities should be, to what extent it should foster treatment on a day-case basis, to what extent it should be integrated with the local university's medical school, and so forth. In the managerialist version of 'competition' that was the 1990s internal market, these variables continued to be set by government policies, not by the competitive process.

2.3. Where the internal market failed

As far as they went, the granting of greater hospital autonomy and the creation of GP fundholding were the moderately successful elements of the internal-market experiment. The attempt to create competition between hospitals, on the other hand, backfired. Propper, Burgess and Green (2004) examined the effect of competition on death rates from acute myocardial infarction. The internal market was introduced England-wide, so there is no obvious control group, but the authors exploited the fact that the intensity of competition differed spatially across England. Purchasers in some localities had a variety of hospitals to choose from, while in others they were confined to a local oligopoly. The authors therefore split regions into different bands according to a measure of 'market' concentration, and looked at the difference-in-differences in death rates, while controlling for differences in patient mix and hospital type. Their findings are sobering:

> We find the impact of competition is to reduce quality. Hospitals located in more competitive areas have higher death rates, controlling for hospital characteristics, actual and potential patient characteristics. . . . [W]hile the estimated impact of competition on quality is small, what it is not is positive. (Propper, Burgess and Green 2004, p. 1267)

A similar study by Propper, Burgess and Gossage (2008) came to more nuanced conclusions, showing that there were positive outcomes as well: hospitals exposed to a greater degree of competition achieved steeper cuts in waiting times, and increased activity levels. But on balance, the impact of hospital competition remained a negative one:

> Using our difference-in-difference approach we find that competition was associated with significantly lower average waiting times and number of persons on waiting lists. Back-of-the-envelope calculations show that this gain does not, however, offset the fall in quality from higher death rates. (Propper, Burgess and Gossage 2008, p. 165)

The most likely explanation was the dearth of information on hospital quality that existed at the time. Commissioners could observe waiting lists, prices, and a hospital's activity levels, but

not standardised mortality rates or infection rates. Faced with competitive pressures, hospitals shifted their efforts from unobservable to observable outcomes. Purchasers were expected to discriminate according to quality, but in the health care environment they found in the 1990s, they were simply not able to do so:

> [O]utcomes measures such as mortality rates were not publicly available . . . until 1999, two years after the competition experiment had ended. . . . Thus purchasers had a strong incentive to negotiate lower prices and/or higher volumes but a much weaker incentive to negotiate (and lower ability to observe) quality improvements or even quality maintenance. (Propper, Burgess and Gossage 2008, p. 142)

At its heart, the internal market was an attempt to run before learning to walk. The preconditions for a functioning market were absent; the ground had not been prepared.

3. The interim period of 'ultra-managerialism'

From 1997 on, the incoming Labour government abandoned most features of the internal market experiment. The purchaser–provider split was nominally retained, but GP fundholding was abolished, effectively with it abolishing competition on the commissioner side. The mission of commissioning bodies – district health authorities, and later the newly formed primary care trusts (PCTs) – was changed to promoting 'cooperation', not competition, between providers (Mays, Dixon and Jones 2011). Health policy in the first term of the Labour government was marked by a return to the aspiration of providing uniform national standards of care within a unified organisation. The 'N' in NHS took centre stage once again. Competition had no place in this approach, both because it was seen as 'divisive' and damaging to an imagined 'public sector ethos' and because competition necessarily entails variation.

This period saw the establishment of what is now the National Institute for Health and Care Excellence (NICE), of the National Service Frameworks (NSF), and of what is now the Care Quality Commission (CQC), all of which were intended to harmonise health provision. NICE provides recommendations on PCT's funding decisions on the basis of cost-effectiveness appraisals. NSF are clinical guidelines which aim to identify and codify medical best practice, and the CQC is an overall regulator and inspector of health care facilities.

From 2000 on, in the wake of the publication of the *NHS Plan* (Department of Health 2000), top-down performance management was intensified. One of the first high-profile measures was the publication of 'star ratings' for NHS trusts. These ratings took account of measures of clinical and financial outcomes, waiting times, patient and staff reviews, and a range of other indicators. Star rating outcomes could have real consequences for a trust's future. The period saw an unprecedented increase in health care spending, and some of the extra funding was channelled through a 'performance fund', access to which was limited to the high achievers in the star ratings. Trusts which did not gain any stars, meanwhile, had to put up with inspections and extensive interference. Senior management staff could even be sacked on this basis.

The rating results were also heavily publicised, and received a great deal of attention. This allowed a level of transparency and public scrutiny which was previously unimaginable, and, at least to a small extent, it probably undermined the deferential attitude towards the NHS. Nigel Crisp, the Chief Executive of the NHS at the time, describes the approach in the following terms:

> [W]e gave the service some major top-down shocks to get it moving – primarily through a policy of publicly 'naming and shaming' the worst performers coupled with a much tougher approach to holding people to account for performance. . . . The NHS generally hated it. (Crisp 2011, p. 58)

The second major policy change of the early 2000s was the introduction of quantitative performance targets. Hospitals were given a schedule outlining by how much they had to reduce waiting times, hospital infection rates, and so forth, in a given time frame. The extent to which hospitals met their targets was a crucial feature of the star ratings, so the two policies were, in practice, tightly interwoven.

The performance management policies of the early 2000s have become very unpopular in the meantime. In the 2010 general election campaign, the opposition Conservative and Liberal Democrat parties pledged to abolish them altogether, and even the Labour government promised to scale them back if re-elected. They have come to stand for an excessive concentration of power at the centre, and for an approach of bureaucratic box-ticking which puts standardised procedures and protocols before common sense.

This critique would have been justified if the alternative to government-imposed discipline had been consumer-imposed discipline, that is, direct accountability to patients. But that alternative was not on offer at the time. Rather, the alternative would have been an arrangement in which providers would not have been held to account by anyone, whether politicians or patients. To put it bluntly: if there is anything worse than a state-managed system, it is a state-managed system in which the state refuses to manage.

The targets/rating system had many severe shortcomings that are well known and well documented (see Bevan and Hamblin 2009 for a review). These shortcomings are inevitable when consumers ultimately have no power over providers. But in so far as a managerialist approach can work at all, the performance management of the early 2000s did work. At the time when the policy was introduced, 80,000 patients had been waiting for hospital admission for more than 15 months. In accident and emergency (A&E) departments, one in five patients waited for more than four hours. The target regime sought to reduce maximum waiting times to 39 weeks in 2005, alongside cutting median waiting time. Ninety-eight per cent of A&E patients were to be treated within four hours, while infection rates were meant to be brought under control at a later stage. All the important targets were met (Crisp 2011, pp. 55–70).

Given the increase in funding, some improvements would have occurred anyway. But comparisons with the smaller nations of the UK, which initially ran a much less stringent performance management regime than England, show that targets and ratings also had an independent effect. A literature review finds:

> Some of this literature has shown reported performance improving in England against the most important targets. All the criticisms of star ratings recognize its undeniable effect, but have also identified six significant general problems: in measuring what matters, selection of targets, nature of measures, aggregation for ranking, gaming and damaging morale. (Bevan and Hamblin 2009, p. 169)

One study looked at changes in waiting times specifically in the English–Welsh border region (Hauck and Street 2007), where socio-economic determinants of hospital performance were, to some extent, automatically controlled for. The authors also took account of the hospitals' overall activity levels and mortality rates, to test whether improvements in waiting lists had simply been achieved by cutting back on the quality and/or quantity of the services provided. They found:

> The English hospitals increased levels of activity, reduced length of stay and undertook proportionately more day case activity over the period. Activity levels remained constant at the Welsh hospital, the proportion of day case activity fell, and proportionately more non-elective patients were admitted. There is no evidence that the English hospitals achieved activity increases by compromising on quality. Mortality rates at the English hospitals remained low or declined further over the period, but the high and rising hospital mortality rates at the North East Wales Trust are cause for concern. (Bevan and Hamblin 2009, p. 288)

Even the NHS's former chief executive concedes that the centralised performance management approach had been taken too far (Crisp 2011, p. 65). But the results show that before the 2000s there had been efficiency reserves in the system which could be exploited even within the given organisational structure. Centralised performance achieved just that.

4. The quasi-market of the 2000s

4.1. Patient choice

From 2002 on, market mechanisms slowly began to creep back in again. An important step was the introduction of patient choice between providers. This started very slowly, initially applying only to selected areas and selected patient groups. In 2002 local choice pilot projects were launched. Patients who could not be treated within six months by the hospital they had been referred to were given the option to switch to a hospital with shorter waiting times. Interestingly, this move was not presented as a means to promote competition between hospitals, but purely as a means to reduce waiting times by allocating patients more efficiently.

Despite the limited scale, the early experience with patient choice was clearly positive. Dawson et al. (2007) estimated the effect of the London Patient Choice Project (LPCP) on waiting times at London hospitals. It is not self-evident that patient choice would reduce waiting times; in a nationalised health system waiting lists are an important rationing tool. It is a theoretical possibility that while a better allocation of patients may initially reduce waiting times, these shorter waiting times could then trigger an increase in demand, pushing waiting times back up again. The authors therefore tested the outcomes by comparing the trend in London waiting times with the trend in the national average, the average for metropolitan areas, and the average for a more specifically selected control group. In order to control for selection bias, the London figures included all London hospitals, even those that did not participate in the LPCP. The authors found:

> The β_1 coefficients indicate the overall difference in waiting times between LPCP hospitals and the control groups. This is significant and negative relative to the Rest of England and matched control groups when using random effects. The size of the coefficients in these two control groups suggest that LPCP waiting times were between 3 and 4 weeks shorter. (Dawson et al. 2007, p. 119)

Choice of an alternative provider after six months was subsequently rolled out nationwide. From 2006 on, it was taken a step further: GPs now had to offer all patients a choice between four or five different providers at the point of referral, and one of these options had to be from the independent sector. In 2008 patient choice at the point of referral was extended to any eligible provider. Choose and Book, an online booking system, and NHS Choices, a website

with information on provider performance, were launched simultaneously. Informed choice was meant to be not just a theoretical possibility but a convenient and accessible option.

That, at least, was the situation on paper. Implementation on the ground lagged behind because many GPs boycotted the reform. In the first year only about a third of patients were actually offered a choice by their GP upon referral, a share which rose to about half over the next few years. Those who were given a choice were usually not given the full range, and private providers in particular were almost never included among those options. In surveys, a majority of GPs expressed negative views of the reforms (Dixon and Robertson 2011, pp. 54–5). So, as a result of resistance from providers, the choices that existed on paper did not fully translate into choices actually experienced by patients. For patients the NHS has been a choice-free environment for most of its history, and many GPs were apparently uncomfortable with attempts to change that.

And yet, in many other sectors, relatively low rates of provider switching are already sufficient to stimulate competition. Contrary to an often-heard critique, markets do not rely on perfectly informed consumers who switch providers all the time. Patient choice did not go as far as intended, but far enough to have some impact.

4.2. The money follows the patient

Patient choice became meaningful only because it was coupled with a wholesale reform of the payment system. Until 2003 NHS hospitals had been paid through annual block contracts. In practice this meant that the main determinant of a hospital's budget in any given year was its budget in the year before. Since the link between a hospital's revenue and its level of activity was weak, hospitals had no incentive to attract patients. This changed in 2003 when the gradual rolling out of an alternative payment system, termed Payment by Results (PbR), began. Payment by Results is a misnomer for a system that should really be called 'Payment by Activity'. It is a prospective payment system under which providers are paid a standardised tariff per case, set on the basis of average cost, with some adjustment for case severity and regional wage/price variation. The basic idea behind PbR is that hospitals should be incentivised to attract more patients, but not to over-treat them. Attracting an additional patient would lead to additional revenue, but a long hospital stay and/or extravagant additional treatments would not.

The introduction of PbR has made the NHS more similar to continental European social insurance systems, although the latter have come from the opposite end. Those systems, as well as the US system, have traditionally run pure fee-for-service reimbursement schemes, where providers are paid for every individual service they perform. While the old British system provided incentives for under-treatment, the fee-for-service systems provide incentives for over-treatment. It would usually pay to keep a patient in hospital for an extra day, recommend an additional test, and so on. Thus, in the continental European context the move towards prospective payment systems represents a move towards greater standardisation; in the UK context it represented a move towards greater differentiation.

PbR initially covered only selected treatments in selected hospitals, with the remainder of hospital revenue still coming through the old block contract system. But it was subsequently expanded to more providers and more procedures. By 2006 about 60 per cent of hospital revenue came from PbR payments (Gaynor, Moreno-Serra and Propper 2011, p. 11). Since

then, however, implementation has stalled. The original intention was that at some point almost all health care spending, not just hospital spending, would be channelled through the PbR system. This has not materialised. In 2010 PbR spending still accounted for only about a quarter of total health spending (Farrar, Yi and Boyle 2011, p. 68).

The third major ingredient in this reform package was the creation of Foundation Trust (FT) status hospitals, which are largely self-governing entities. Hospitals could apply for FT status when they met specified standards of clinical and financial performance ('earned autonomy'). The first conversions to FT status occurred in 2004, and by 2010 131 NHS hospitals had become FTs (Allen and Jones 2011, p. 25). But, again, the process stalled. The original intention was that virtually all hospitals would eventually acquire FT status, which has not been achieved.

Even though Labour's health policy developed in a very haphazard way, this package of market-oriented reforms shows a remarkable degree of internal consistency. The combination of patient choice and PbR created a system in which the money followed the patient: patients could now choose providers, and the choices they made had a real financial impact on those providers. For the first time since 1948, the revenue of health care providers would, to a considerable extent, depend on the free choices of patients, giving providers a good reason to be responsive to those patients' needs. It is only in this context that the introduction of FT status also became sensible. Now that providers were more directly accountable to their patients (and potential patients), government interference with their day-to-day operations became less necessary. The discipline of the quasi-market could replace government-imposed discipline. Competition made greater autonomy possible and, indeed, necessary. If providers were to cope with competitive pressures, they also had to be given the leeway to respond to those pressures. They had to be given the managerial autonomy to reshape their organisations accordingly.

And despite the weaknesses highlighted above, the quasi-market reforms of the 2000s were also clearly superior to the internal market reforms of the 1990s in that they had been preceded by improvements in the availability of information on provider quality. This time round, the seeds of market reforms were sown on well-prepared ground. The reforms of the 2000s also went beyond those of the 1990s in that the choice of hospital was given to patients, not their GPs (at least in theory). Of course, for the vast majority of patients GPs still are, and will remain, the most important source of information and advice when making that choice. This is part of their proper role. But the choice is ultimately the patient's.

4.3. The empirical evidence

A number of studies have investigated the impact of choice-driven competition on the quality and efficiency of hospital care (for a shorter review, see Bevan and Skellern 2011). There is no control group as such, as the reforms progressed at a uniform speed England-wide. But the intensity of competition differed across the country. Almost all English patients have access to at least two hospitals within a half-hour travel radius, but areas differ a lot in the number of potential competitors and the market share of the largest provider(s). Studies on the effect of competition therefore map the degree of potential competition that exists in different parts of England, using standard measures of industry concentration and different specifications of market size. They thereby split England into a large number of overlapping hospital markets

with varying degrees of competitiveness. They then attempt to control for factors that are thought to be associated with competition without being causally related to it (alongside the more obvious confounding factors like differences in case mix and case severity).

The least competitive areas are then treated as a quasi-control group. The time trend in their outcomes is interpreted to be the closest proxy of the time trend in England in the absence of reform. Improvements in the least competitive areas can be interpreted as improvements that would have occurred anyway, but, if the more competitive areas experience improvements over and above that baseline, those additional improvements can be ascribed to competition.

In this way, Bloom et al. (2010) studied the relationship between the intensity of competition and the quality of hospital care, approximated by mortality rates from acute myocardial infarction (AMI) and emergency surgery. They found that

> hospitals facing more competition have significantly fewer deaths following emergency AMI admissions. . . . [T]here appears to be a causal effect whereby adding one extra hospital reduces death rates by 1.83 percentage points. (Bloom et al. 2010, p. 14)

Competition also lowers the death rate from emergency surgery.

Gaynor, Moreno-Serra and Propper (2011) estimated the difference-in-differences in mortality rates from AMI, as well as all-causes mortality rates, again comparing hospitals facing varying degrees of competition. Their results show that:

> . . . higher market concentration (a larger HHI^2) leads to lower quality. A 10% increase in the HHI leads to an increase of 2.91% in the AMI death rate. . . . The estimate [for the all-cause mortality rate] again shows a significant relationship between quality and market concentration. The magnitude is smaller than that for AMI but precisely estimated. (Gaynor, Moreno-Serra and Propper 2011, p. 20)

The authors showed that this divergence is not a continuation of a previously existing trend, but a new trend that started when competition became effective. And while the percentage point differences in death rates are modest, for a relatively frequent condition like AMI they still translate into noticeable numbers of lives saved: 'This amounts to . . . a little over 8 fewer AMI deaths annually per hospital, or approximately 1,000 fewer total deaths per year over all 135 hospitals in our sample' (p. 21).

In a similar model which also uses the difference-in-differences in AMI death rate as a proxy for hospital quality, Cooper et al (2011, p. 244) found:

> 30-day AMI mortality fell 0.31 percentage points faster per year after the reforms for patients treated in more competitive markets . . . Framed differently, the shift from a market with two equally sized providers to one with four equally sized providers after the reforms would have resulted in a 0.39 percentage point faster reduction in AMI mortality per year from 2006 onwards.

They, too, ruled out the possibility that this was merely a continuation of a pre-existing trend, or an artefact of how 'competitiveness' was measured:

> An essential observation . . . is that the pre-policy trend in AMI mortality in areas with uncompetitive market structures is not statistically different from the trend in markets with competitive structures

> once we control for patient characteristics . . . Our findings remain consistent and significant across the seven different measures of market structure. (pp. 244–5)

Models of this type have also been used to study the relationship between competition and hospital efficiency. The already mentioned paper by Gaynor et al. (2011) used average length of stay (ALOS) as a proxy for efficiency, alongside information on hospital expenditure per activity. The authors found that hospitals in more competitive markets have recorded greater productivity improvements:

> The estimated coefficient implies that a 10% fall in a hospital's HHI on average results in a 2.3% fall in length-of-stay. . . . Taken together, the findings for quality . . . and resource utilization . . . suggest that hospitals facing more competitive pressure were able to find ways to marshal resources more efficiently to produce better patient outcomes. (p. 22)

Cooper et al. (2012) used a similar model to test the same relationship, and in addition tested whether ALOS reductions represent genuine efficiency improvements or whether they have been achieved by discharging patients sooner than is clinically appropriate ('sicker and quicker'). They did this by splitting ALOS into a pre-surgery component and a post-surgery component, arguing that the former (the time from a patient's arrival at the hospital to the commencement of the procedure) can only be shortened by genuine improvements in the hospital's internal workflow. They found that, measured in this way, competition leads to efficiency improvements:

> . . . a one standard deviation decrease in market concentration pre-reform was associated with a reduction in overall LOS [length of stay] of between 2% and 6% relative to the mean LOS over that period. . . . Framed differently, the addition of one hospital to a hospital market lowered the LOS for patients treated in that area by approximately 0.4 days. (Cooper et al. 2012, p. 18)

However, the authors also found that the inclusion of private hospitals has a negative effect on the productivity of nearby public hospitals. Their explanation is that the latter probably attract patients who are generally healthier, but in ways which the control variables fail to register (and the PbR severity adjustments fail to fully compensate for).

In short, there is good evidence that the market reforms of the 2000s have increased quality and efficiency in the NHS. Some studies have also taken a closer look at the transmission mechanisms behind these results, exploring not just *whether* but also *how* the reforms produced those results.

One part of the answer is that once they were able exercise choice in a meaningful way, patients became more discriminating and quality-conscious. Gaynor, Moreno-Serra and Propper (2011, p. 19) argue:

> If patients became more responsive to quality post-policy we should see better hospitals (those in the bottom quartile of the mortality distribution) attracting more patients relative to worse hospitals (those in the top quartile). That is exactly what the data show . . . [T]he share of patients bypassing their nearest hospital increased for better hospitals while it clearly decreased for worse hospitals. This provides reassurance that there is a patient response to quality and that it increased during the reform.

A paper by Gaynor, Propper and Seiler (2012) examined patient behaviour in greater detail. For patients undergoing coronary artery bypass graft surgery, they estimated the elasticity of demand with regard to quality, with quality being approximated by mortality rates. In other words, they estimated whether hospitals that record low (high) mortality rates experience an increase (decrease) in demand in subsequent years. They found that, before the introduction of choice, when patient demand was mediated through GPs' decisions, the elasticity of demand was indistinguishable from zero. After the introduction of patient choice, it fell to minus 0.12: patients did discriminate against underperforming hospitals. The authors also estimated how far a hospital's mortality rate affects its market share in subsequent years. Again, in the pre-reform period, there was little connection between these two variables. An increase in the mortality rate by one standard deviation would reduce a hospital's market share by only 0.36 per cent. After the reform, though, the same increase in mortality would be punished with a 4.9 per cent loss in market share (p. 24).

The data contradict the widespread notion that health care is not amenable to choice because of its complexity. Rather, people generally seem to make quite reasonable choices. For example, sicker patients were more responsive to quality differences than healthier ones, as one would expect given that more is at stake for them. Income, on the other hand, was a poor predictor of responsiveness. Fears that the well-educated middle classes would choose the best hospitals and leave the less well-off behind did not materialise.

These findings are not incompatible with the fact that, according to their own account, few patients explicitly study clinical data before choosing a hospital even though this is possible through the NHS Choices website. Information on hospital quality, or close correlates, can also be disseminated through more informal channels.

Looking also at the transmission channels through which the market reforms have affected quality, Bloom et al. (2010) inspected what happens inside hospitals, again differentiating them by the extent to which they are exposed to competition. They studied the relationship between competition and an 'index of management quality', which is arguably a misnomer: their index does not measure management quality, which is an outcome, but rather the extent to which an organisation's senior staff *attempt* to manage the organisation sensibly. It measures the extent to which formalised procedures of quality control, monitoring, reporting, accountability, and so on, are in place. Still, the authors showed that, when applied to other sectors, their index correlates with desirable outcome measures, so it does seem to have some predictive power. This is true in health care as well. The authors showed that hospitals that are 'better managed' according to their indicator also record lower mortality rates, shorter waiting lists, lower MRSA infection rates, higher operating margins, and higher levels of job satisfaction among employees (Bloom et al. 2010, pp. 11–12, 23). They also found that management quality is, among other factors, driven by competition. Government targets may, to some extent, have incentivised hospitals to cut corners and game the system, but competition has incentivised them to optimise internal procedures, thus driving genuine improvements in quality and efficiency.

Finally, it is insightful to compare the evolution of the English NHS with that of its counterparts in Scotland, Wales and Northern Ireland. The smaller UK nations are not control groups as such, since they differ from England on many dimensions other than health care. But they have experienced the same or larger increases in funding; and they have, after some delay, introduced similar performance targets. What distinguishes them most clearly from England is that they have not introduced the market-oriented reforms discussed above, or introduced only

isolated elements of them. The results diverge sharply. The smaller UK nations, especially Scotland, have recorded higher per capita levels of health care spending. Proportionally, they employ larger numbers of hospital, medical, dental, nursing, midwifery, health visiting, hospital management and support staff. They record higher numbers of hospital beds and inpatient admissions. And yet they have longer waiting times for inpatient and outpatient appointments, and longer ambulance response times. They do worse on outcome measures across the board, and differences in efficiency have to explain a large share of that. In England, many more hospital patients are treated as day cases, and activity levels are far higher when expressed in per staff terms (Connolly, Bevan and Mays 2010). England's better health outcomes may have many other determinants that lie outside the health system's reach, but when it comes to efficiency differences there is no other plausible candidate in sight.

In short, the reforms of the 2000s appear to have greatly improved the English NHS in a number of ways. For supporters of market-oriented reform, it is tempting to read the literature as a vindication of their position, but that interpretation would probably overstretch the literature. A number of limitations remain. First, the evidence thus far available rests on a relatively narrow range of outcome and quality measures. The authors concerned justify this by arguing that AMI survival rates in particular are positively correlated to outcome measures across the board.[3] But this claim is not uncontroversial: most hospitals are not simply 'good' or 'bad'; rather, there is great within-hospital variation in terms of quality (Bevan and Skellern 2011). The evidence also rests on a relatively short time period, so it is as yet unclear whether the improvements that have been made can be sustained. Many other features of the health care landscape are changing as well, so it is not clear whether the results that have been recorded under the 'old' conditions can be reproduced in the health care environment that is currently being developed. It is also true that many other potential effects of competition have not yet been tested. Critics might argue that the reforms have turned former partners into competitors, and thereby reduced their willingness to cooperate and share expertise. So far, this possibility cannot be ruled out.

But even if future empirical studies could overcome all those limitations, and turn tentative findings into definitive ones, they would still not represent a vindication of competition in health care per se. They would at best vindicate a very specific version of a competitive process, namely competition subject to centrally determined prices and to centrally collected information on outcomes and quality. One cannot necessarily extrapolate from this specific set-up to other versions of competition; in particular, one cannot extrapolate to price competition or to a set-up in which the dissemination of information is left to providers. Still, subject to those constraints, the evidence on patient choice and competition that is so far available is predominantly favourable.

5. Conclusion and outlook

While the first attempt to introduce market mechanisms into the NHS did not succeed, the second attempt did. The main difference has to be that the internal market of the 1990s was a premature birth. Competition was introduced in the context of a dearth of information, in which commissioners were unable to assess provider quality. Even though other aspects of the internal market – GP fundholding and the creation of NHS trusts – were successful, hospital competition, under those circumstances, led to perverse results.

The subsequent period of managerialism may have taken centralised performance management through targets and ratings too far, but one of the positive legacies of that era is that the amount of information on provider outcomes increased vastly. This turned out to be a boon when competitive mechanisms were reintroduced in the mid-2000s, because this time the preconditions for a functioning market were in place. Combined with patient choice, greater autonomy for providers, and a payment system in which funding followed patients, this turned out to be a potent and coherent reform package. The reforms drove up quality, productivity and other measures of performance.

These reforms can be built on. The most obvious first step is to finish the job. This would mean converting all hospitals into Foundation Trusts, channelling all health care funding through the Payment by Results system, and requiring all GPs to inform their patients about their right to choose providers. But to take competition to a higher level, the market has to be moved from static competition to dynamic competition – a market with entries and exits. A strict no-bail-out clause for failing providers, alongside the right of independent sector organisations to take over insolvent providers, would go a long way towards this. Hospital bankruptcies would then become a normal occurrence – although 'bankruptcy', in this case, would probably not mean that a facility shuts down, but that it would find itself under new management.

Competition should also be extended to the commissioning side and to primary care. This would mean an abolition of the catchment area system, and thus a breaking of the link between commissioning, primary care, and place of residence. Patients would be able to choose any primary care trust or clinical commissioning group as well as any GP they see fit, wherever they are based. This would effectively convert the British single-payer model into a multiple-payer, social-insurance model. This model is not per se superior to a NHS model, but it does offer the potential benefit of making commissioners/insurers more responsive to patient needs, and shape health care provision in their favour (Higgins 2007; Hussey and Anderson 2003).

This could then be coupled with opening up the commissioning side. Just as there are independent sector providers, there could be independent sector commissioners as well. These could be private companies, civil society organisations or communities of interest. Trade unions, professional associations and religious organisations would be obvious candidates, and so would be patient advocacy groups that could specialise in the conditions that are most relevant to their members. Risk structure compensation would ensure a level playing field between commissioners.

These commissioners should also be allowed to merge and integrate with provider organisations at various levels, which, in turn, should be allowed to merge and integrate with one another. A variety of different patterns of vertical and horizontal integration could then be tried and tested.

Commissioners should also be given the freedom to experiment with different financial incentive schemes encouraging patients to economise on health care consumption. This can include optional co-payments and deductibles in return for tax rebates, as well as a self-commitment to use lower tiers of health care provision first. Such schemes are already established practice in Switzerland,

The establishment of a quasi-market has benefited NHS patients. The logical next step is to move on from a managerialist understanding of competition towards a richer, more comprehensive understanding. In the current system, competition takes place within politically

determined structures of health care delivery. It should be replaced by a system in which those delivery structures are themselves determined by the competitive process.

Notes

1. 'Difference-in-differences' methodology is a widely used econometric approach which attempts to simulate a controlled experiment. It estimates the effect of a 'treatment' (in this case a change in fundholding) on an outcome (in this case waiting times) by comparing the average change over time in the outcome variable for the treatment group with the average change over time for the 'control group' (those whose GPs did not become fundholders).
2. HHI = the Herfindahl–Hirschman Index, a measure of market concentration, with lower values indicating more intense competition.
3. For example, Gaynor, Moreno-Serra and Propper (2011, p. 15) argue that 'AMI mortality is treated as an indicator of overall quality in the hospital for a number of reasons. First, the infrastructure used to treat AMI is common to other hospital services, making it a good general marker of hospital quality . . . Second, AMI admissions are reasonably high volume and mortality is a fairly common outcome so variability in the rates is less of an issue than for other treatments. Third, as all patients with a recognized AMI are admitted there is little scope for selection bias.'

References

Allen, P. and L. Jones (2011) 'Increasing the Diversity of Health Care Providers', in Dixon, Mays and Jones (2011).

Bevan, G. and R. Hamblin (2009) 'Hitting and Missing Targets by Ambulance Services for Emergency Calls: Effects of Different Systems of Performance Measurement within the UK', *Journal of the Royal Statistical Society* 172(1), 161–90.

Bevan, G. and M. Skellern (2011) 'Does Competition Between Hospitals Improve Clinical Quality? A Review of Evidence from Two Eras of Competition in the English NHS', *British Medical Journal* 343, d6470. Available at http://eprints.lse.ac.uk/40065/ (accessed 9 December 2014).

Bloom, N., C. Propper, S. Seiler and J. Van Reenen (2010) The Impact of Competition on Management Quality: Evidence from Public Hospitals. NBER Working Paper 16032. Cambridge, MA: National Bureau of Economic Research.

Connolly, S., G. Bevan and N. Mays (2010) *Funding and Performance of Healthcare Systems in the Four Countries of the UK Before and After Devolution: A Longitudinal Analysis of the Four Countries, 1996/97, 2002/03 and 2006/07, Supplemented by Cross-Sectional Regional Analysis of England, 2006/07*. London: The Nuffield Trust.

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

Cooper, Z., S. Gibbons, S. Jones and A. McGuire (2012) Does Competition Improve Public Hospitals' Efficiency? Evidence from a Quasi-Experiment in the English National Health Service. CEP Discussion Paper No. 1125. London: Centre for Economic Performance, London School of Economics and Political Science.

Crisp, N. (2011) *24 Hours to Save the NHS: The Chief Executive's Account of Reform 2000–2006*. Oxford: Oxford University Press.

Dawson, D., H. Gravelle, R. Jacobs, S. Martin and P. Smith (2007) 'The Effect of Expanding Patient Choice of Provider on Waiting Times: Evidence From a Policy Experiment', *Health Economics* 16, 113–28.

Department of Health (2000) *The NHS Plan: A Plan for Investment, A Plan for Reform* (Cm 4818-I). London: The Stationery Office.

Dixon, A., N. Mays and L. Jones (eds) (2011) *Understanding New Labour's Market Reforms of the English NHS*. London: The King's Fund.

Dixon, A. and R. Robertson (2011) 'Patient Choice of Hospital', in Dixon, Mays and Jones (2011).

Farrar, S., D. Yi and S. Boyle (2011) 'Payment by Results', in Dixon, Mays and Jones (2011).

Gaynor, M., R. Moreno-Serra and C. Propper (2011) Death by Market Power. Reform, Competition and Patient Outcomes in the National Health Service. Working Paper No. 10/242. Bristol: Centre for Market and Public Organisation, University of Bristol.
Gaynor, M., C. Propper and S. Seiler (2012) Free to Choose? Reform and Demand Response in the English National Health Service. CEP Discussion Paper No. 1179. London: Centre for Economic Performance, London School of Economics and Political Science.
Hauck, K. and A. Street (2007) 'Do Targets Matter? A Comparison of English and Welsh National Health Priorities', *Health Economics* 16, 275–90.
Higgins, J. (2007) 'Health Policy: A New Look at NHS Commissioning', *British Medical Journal* 334(7583), 22–4.
Hussey, P. and G. Anderson (2003) 'A Comparison of Single- and Multi-payer Health Insurance Systems and Options for Reform', *Health Policy* 66, 215–28.
Mays, N., A. Dixon and L. Jones (2011) 'Return to the Market: Objectives and Evolution of New Labour's Market Reforms', in Dixon, Mays and Jones (2011).
Propper, C., S. Burgess and D. Gossage (2008) 'Competition and Quality: Evidence from the NHS Internal Market 1991–9', *The Economic Journal* 118, 138–70.
Propper, C., S. Burgess and K. Green (2004) 'Does Competition Between Hospitals Improve the Quality of Care? Hospital Death Rates and the NHS Internal Market', *Journal of Public Economics* 88, 1247–72.
Propper, C., B. Croxson and A. Shearer (2002) 'Waiting Times for Hospital Admissions: The Impact of GP Fundholding', *Journal of Health Economics* 21, 227–52.
Söderlund, N., I. Csaba, A. Gray, R. Milne and J. Raftery (1997) 'Impact of the NHS Reforms on English Hospital Productivity: An Analysis of the First Three Years', *British Medical Journal* 315(7116), 1126–9.
Taylor, R. (2013) *God Bless the NHS: The Truth Behind the Current Crisis*. London: Faber & Faber.