Centre for Economic and Financial Research at New Economic

The adverse effect of incentives regulation in healthcare: a comparative analysis with the U.S. and Japanese hospital data

Galina Besstremyannaya

Working Paper No 218

CEFIR /NES
Working Paper series

The adverse effects of incentives regulation in health care: a comparative analysis with the U.S. and Japanese hospital data

Galina Besstremyannaya *

Revised October 9, 2015

Abstract

The paper analyzes the effect of incentives regulation, when the yardstick competition approach is supplemented with a performance tax on providers. In an application to prospective payments in health care in the U.S. and Japan, we show differential effects of value-based purchasing, when price-setting is related to benchmark values of quality measures or length-of-stay. The predictions of our theoretical model, as well as empirical results offer persuasive evidence that unintended effects appear for best-performing hospitals. Patient experience/clinical-process-of-care measures significantly decrease in the top percentiles of the U.S. hospitals owing to the reform. Similarly, length of stay significantly increases for most diagnosis-related groups at Japanese hospitals in percentiles with the lowest length of stay. A natural experiment aimed at best-practice rate-setting diminishes the undesired effects of the reform.

JEL Classification Codes: C22, C23, D21, D22, I18

Keywords: quantile regressions, prospective payment, hospital financing

*Lead researcher, CEFIR at New Economic School

1 Introduction

Public contracting under asymmetric information may be viewed as a classic example of an agency problem, where a social optimum in terms of minimizing provider costs can be achieved with nonlinear pricing (Joskow and Rose (1989)). A particularly notable implementation is yardstick competition – setting the cost of comparable firms as the benchmark for a given firm (Shleifer (1985)). However, providers face a number of objectives, such as issues related to costs, quantity and quality. This context of a multi-task agency problem may result in a trade-off between quality and cost efficiency, especially if demand does not respond to quality (Chalkley and Malcomson (1998), Holmstrom and Milgrom (1991)). A solution to the problem has been found in the mechanism of incentives regulation with taxation and redistribution, stemming from pay-for-performance in managerial economics. Applied to remuneration in various private and public industries (e.g. telecommunications, electricity, federal government, health and education), a pay-for-performance schedule rewards providers who reach a target value of a certain performance indicator and may punish providers below the target. Alternatively, pay-for-performance may stimulate providers in the upper tiers of the indicator distribution and decrease payments to providers in the lower tiers. Bonus payment may be granted for improvement over time.

Various forms of pay-for-performance are particularly prevalent in health care, which is a classic case of an industry with an asymmetric information and physician agency problem. While a number of works in the literature concentrate on the overall effect of the reimbursement schedule and heterogeneous impact across providers, such approaches have substantial shortcomings. Firstly, theoretical models often produce ambiguous predictions, depending on the concavity conditions and values of the parameters (Besstremyannaya and Shapiro (2012), Miraldo et al. (2011), Grabowski et al. (2011), Christianson and Conrad (2011)). Secondly, there are issues about the internal validity of empirical estimates, since analyses commonly do not account for endogeneity of participation in the reform and endogenous price-setting, related to the empirical distribution. Moreover, the studies generally employ data for composite measures, aggregate diagnosis groups or selected diagnoses. Finally, despite theoretical arguments advocating the desirability of best-practice price-setting over conventional benchmarking, there is, to the best of our knowledge, no study evaluating the effect of the corresponding change.

The present paper contributes to the literature, overcoming the above shortcomings as follows. We propose a theoretical model, which forecasts the adverse effects of performance-based reimbursement for hospitals with the best target indicators. Two versions of the model apply respectively to quality and length-of-stay performance, as implemented within recent hospital financing reforms in the U.S. and Japan.

The novelty of our empirical approach is threefold. Firstly, instead of statistical analysis, ordinary-least squares models with dynamic panel data or simple quantile regression framework, we use dynamic panel data quantile regressions. We use a "habit-formation" model (e.g. autocorrelation, resulting in endogeneity, both for the U.S. and Japan) and account for endogenous participation in Japanese reform. We extend the Canay (2011) methodology for two-step estimation of panel data quantile regressions with endogenous variables. Adding an independence of disturbance term from lagged endogenous covariates, at the first step, we consistently estimate fixed effects using the Arellano and Bover (1995)/Blundell and Bond (1998) estimator, with robust variance-covariance matrix (Windmeijer (2005)). At the second step we modify the Chernozhukov and Hansen (2004) grid-search procedure for instrumental variable estimation of a two-dimensional vector of endogenous variables and a large number of instruments. The Wooldridge (2007) correction for serial correlation in the random effects panel data model is extended in this paper for the instrumental variable regression.

Secondly, we use similar empirical framework for estimates with the latest nationwide longitudinal data

for the two countries, taking hospital /diagnosis-group level administrative panel data on a changeover to performance-based remuneration in the U.S. and Japan (Hospital Compare data, Medicare Impact files and Finale Rules for 4048 hospitals in fiscal years 2008–2013, as well as Medicare provider utilization and payment data in 2011–2013;[1] Japan's Ministry of Health, Labor and Welfare data base on 1849 hospitals in Jul 2005 – Mar 2014 and Ministry of Internal Affairs and Comunications data on all municipal and regional hospitals in April 1999– March 2014).[2]. The non-rejection of our theoretical hypotheses in this context may justify the external validity of the approach.

Thirdly, the previous analyses with pay-for-performance in the U.S. used the data for prototypes of value-based purchasing and concentrated on composite measures (Ryan et al. (2012), Werner and Dudley (2012), Borah et al. (2012), Werner et al. (2011), Lindenauer et al. (2007)). Contrary to using the preliminary data on the U.S. pilot implementation, we employ nationwide databases and concentrate on each quality measure. We discover that the effect of value-based purchasing varies for different quality measures. As regards the Japanese length-of-stay reimbursement, this paper is the first to analyze nationwide data at the diagnosis group level and to study the effect of the change in the payment schedule. Our estimations demonstrate that although the heterogeneity in the reform effect is similar across major diagnostic categories, it differs for various diagnoses within each category.

Our results offer persuasive evidence supporting the adverse effects of pricing on quality or length-of-stay performance. Measures of patient experience of care significantly decrease in the top percentiles of the U.S. hospitals. Similarly, average length of stay significantly increases for Japanese hospitals in percentiles with the lowest length of stay. A natural experiment with a step towards the best-practice rate setting in Japan diminishes the adverse effects of the reform.

The remainder of this paper is structured as follows. Section 2 reviews the literature on incentives regulation and its health care applications. Section 3 explains pay-for-performance systems in the U.S. and Japan. Section 4 provides theoretical model, predicting heterogeneous effects of performance-based reimbursement, and describes econometric methodology for estimating dynamic panel data models with endogeneity. Section 5 outlines the data for each country. Section 6 presents the results of the estimations, and the discussion about price-setting is given in Section 7.

2 Related literature

The origins of incentives regulation under asymmetric information may be found in the approach by Baron and Myerson (1982) and the yardstick competition model by Shleifer (1985), which aims to set a benchmark for evaluating the potential for a regulated monopolistic firm. The model (often called a fixed-price contract) establishes the price for each firm dependent on the costs of similar firms and independent of the firm's own price.

In an application to health care, yardstick competition requires the identification of a hospital's products and determination of a reasonable cost for each product. This is accomplished with the help of diagnosis related groups (DRGs), developed in the 1960s by the Yale University Center for Health Studies as a system for describing hospital production (Fetter and Freeman (1986)). DRGs classify patients into a restricted number of medically justified groups, with a statistically stable distribution of resource consumption within each group (Thompson et al. (1979)). This classification is a core part of a prospective payment system (PPS) – a method of reimbursement that provides fixed payments for a patient with a given DRG. Piloted in New Jersey in the 1980s and then applied to all Medicare hospitals in the United States, this approach

[1] Dec 2014/Jun 2015 updates
[2] Sep 2014/May 2015 updates

has been adopted in most health care systems. It may be noted that such average cost pricing is a version of yardstick competition, when lump-sum transfers are unavailable.

Laffont and Tirole (1986) extend the approach to the case when a firm's cost-reducing efforts are not observed. The authors propose a two-part tariff, which is the sum of a fixed price (a lump-sum transfer) and a fraction of actually incurred costs. The purpose of the tariff is to share risks owing to uncertainty about the firm's costs.

If quality and output are independent objectives, then quality may be regarded as an additional output in the framework of the multi-product firm (Laffont and Tirole (1990)) and the same model may be used. However, quality and output are likely to be dependent. Therefore, the incentives for quality enhancement and cost-reducing efforts should be analyzed in their totality as interrelated objectives of a multi-task agency problem (Holmstrom and Milgrom (1991)). In this regard, Laffont and Tirole (1993) investigate the influence of quality on the power of incentive schemes and discover differential results depending whether quality and quantity are net complements or net substitutes. Ma (1994) shows that prospective payment leads to efficient levels of costs and quality when these are the only two objectives of a hospital; while incentive trade-offs arise in the presence of other objectives.

A solution for dealing with such trade-offs may be discovered in incentives regulation, related to performance-based reimbursement. It dates back to the early 1980s when various performance targets were employed for enhancing the quality of natural monopolies and telecommunications (Kridel et al. (1996), Joskow and Schmalensee (1986)). In the health sector, the history of nationwide implementation of pay-for-performance (also called "payment by results") starts with 136 measures for family practices in the U.K. These targets were established in 2004 and covered patient experience of care, management of chronic diseases and practice organization (Campbell et al. (2009)). Overall, health care attracts major attention in terms of performance-based reimbursement, owing to the large share of public expenditures and the presence of welfare issues, demanding regulation (Chalkley and Malcomson (2000)).

It may be noted that numerous private and public programs linking quality and reimbursement in health care existed in the U.S. in the early 2000s, mostly at employer or state level (Ryan and Blustein (2011), Damberg et al. (2009), Pearson et al. (2008), Rosenthal (2008), Damberg et al. (2005), Rosenthal et al. (2004)). Later the Hospital Quality Incentive Demonstration (HQID) used 33 quality measures of the clinical process of care for a pilot with quality-performance reimbursement to Medicare hospitals. The success of the pilot in terms of average enhancement of hospital quality has resulted in the nationwide introduction of pay-for-performance within a prospective payment system – a value-based purchasing reform, started in 2013. Regarding length-of-stay performance, Belgium and Japan adjust prospective tariffs according to a hospital's position relative to the values of percentile points in the nationwide distribution. The system leads to decrease of the average length of hospital stay at the country level.

Concerning theoretical analysis on incentives regulation in health care, Chalkley and Malcomson (1998) devise a price schedule, aimed at decreasing costs and sustaining quality levels, when patient demand is quality inelastic. Their types of contract depend on whether a hospital is interested solely in its profits or is benevolent to patients, at least to a certain extent. Miraldo et al. (2011) concentrate on optimal price schedule within prospective payment, when a hospital's costs are fully observable and no lump-sum transfers are allowed. The authors show that the optimal price depends on the interrelation between a hospital's heterogeneity parameter and quality. Grabowski et al. (2011) uses the general framework of Ma (1994) to study prospective payments, which continuously depend on per diem intensity of care. The paper demonstrates that the effects of payment generosity on such a performance indicator as length of stay is ambiguous. In a related paper, Besstremyannaya and Shapiro (2012) regard a hospital's length of stay as a function of intensity of care, and argue that the effects of performance-based per diem schedule depend on

the concavity of the length-of-stay function.

The closest theoretical paper, related to ours, is Grabowski et al. (2011). We build upon the framework of the prospective price, related to treatment intensity, and use the volume of service, calculated as the number of patient-days. Our contribution consists in adding quality as an additional argument and analytical solving the model for the cases with performance-based payment schedules in the U.S. and Japan.

As for the empirical analysis, the theory and inference for an instrumental variable approach, allowing consistent estimation of quantile regression with endogenous covariates, as well as a practical implementation may be found in a cross-sectional model of Chernozhukov and Hansen (2008), Chernozhukov and Hansen (2006), Chernozhukov and Hansen (2004). Galvao (2011) shows consistency of Chernozhukov and Hansen (2004) approach in case of panel data models with endogenous variables, using an example of AR(1) dynamic panel data model. Parente and Santos Silva (2015) and Wooldridge (2007) proposes a correction for serial correlation in the random effects panel data model, which we extend in this paper to the instrumental variable regression. Concerning panel data quantile regression with quantile-independent fixed effects, Canay (2011) proposes a computationally simple two-step estimator, which first, consistently estimates fixed effects under the assumption that they are "locational shifts", computes fitted value of the dependent variable (subtracting the fitted value of "locational shifts"), and second, applies panel data quantile regression methodology to the fitted value of the original dependent variable.

The novelty of our approach is the extension of quantile regression methodology for fixed effects dynamic panel data models with endogeneity. We modify the Canay (2011) approach for two-step estimation of panel data quantile regressions with endogenous variables and extend the Chernozhukov and Hansen (2004) instrumental variable estimations. The empirical estimates with our dynamic panel data quantile regression model incorporate the issues of "habit-formation" and endogenous reform participation.

The findings about the results of pay-for-performance generally show that this reimbursement mechanism enhances the mean level of performance (Eijkenaar et al. (2013), Houle et al. (2012), Moreno-Serra and Wagstaff (2010)). However, the observed mean effect hinders differential responses by under-performing and over-performing providers. According to the theoretical models and empirical evidence about evolution of quality and length-of-stay in the U.S., U.K. and Japan, providers already above the target may not have enough incentives for improvement (Nawata and Kawabuchi (2013), Ryan et al. (2012), Besstremyannaya and Shapiro (2012), Werner et al. (2011), Miraldo et al. (2011), Grabowski et al. (2011), Mannion et al. (2008), Doran et al. (2008), Lindenauer et al. (2007), Rosenthal et al. (2005)).[3]

Our empirical findings are close to the results of a few papers on quality deterioration of composite measures in the best-performing U.S. hospitals and increase of length of stay at the aggregate level of major diagnostic categories in the Japanese hospitals (Nawata and Kawabuchi (2013), Ryan et al. (2012), Besstremyannaya and Shapiro (2012), Werner et al. (2011), Miraldo et al. (2011), Grabowski et al. (2011), Doran et al. (2008), Lindenauer et al. (2007)). The novelty of our analysis is the concentration on each quality measure and each diagnosis-procedure combination, which are the basic units in the health care industry. We discover a variation of the effect across quality measures and diagnosis-groups, which may be related to skewness of the nationwide distribution and medical issues, concerning resource consumption.

[3]Additionally, some literature focuses on inequality effects for patients, grouped according to the value of the performance target (e.g. length-of-stay, Sood et al. (2008), McKnight (2006), Ellis and McGuire (1996)).

3 Performance-based reimbursement

3.1 Value-based purchasing in the U.S.

Value-based purchasing (VBP) applies to all discharges within the inpatient prospective payment system for Medicare hospitals starting October 2012. The reform decreases Medicare's payment to each hospital by a factor α and redistributes the accumulated fund. A hospital's rewards are based on a linear exchange function, translating *total performance score* into payments, so that the adjustment coefficient γ_i for each hospital i become:

$$\gamma_i = 1 + \big(s\frac{tps_i}{100} - 1\big) \cdot \alpha, \tag{1}$$

where tps_i is hospital's *total performance score* ($0 \leq tps_i \leq 100$), s is the slope of a linear exchange function, which is set at the level 1.93621799 for FY 2013 and 2.0961880387 for FY 2014, $\alpha = 0.01$ in FY 2013 and is increased by 0.0025 percentage points in 2014-2016 to reach 0.02 from 2017 onwards. Intuitively, if $s = 2$ then all hospitals performing below the national mean of *tps* are financially punished, as their $\gamma_i < 1$.[4]

The *total performance score* is computed on the basis of scores for the measures of *clinical process of care* domain and *patient experience of care* domain.[5] The score for each clinical process of care measure is the percent of patient cases, for which the corresponding clinical requirement was satisfied. Regarding patient experience of care measures, the score is the percent of discharged patients who gave the most positive ("top-box") response to the corresponding question (Table 6).

For each hospital i and each measure m in any domain *achievements points* a_i^m ($0 \leq a_i^m \leq 10$) are calculated as follows:

$$a_i^m = \begin{cases} 10, & \text{if } y_i^m \geq m_b \\ \text{Round}\left[\dfrac{9(y_i^m - m_a)}{m_b - m_a} + 0.5\right], & \text{if } m_a \leq y_i^m < m_b \\ 0, & \text{if } y_i^m < m_a \end{cases}$$

where y_i^m is the score for measure m for hospital i in the current period, m_b is benchmark, m_a is achievement threshold for measure m. In other words, a hospital receives a maximum value of 10 achievement points if its quality score is above the benchmark, a minimum value of 0 points if the score is below the threshold; and a value between 0 and 10 (rounded to the closest integer), which positively depends on a hospital's distance from the threshold.

Improvement points p_i^m ($0 \leq p_i^m \leq 9$) are computed as the difference between a hospital's score in the current period and the baseline period, normalized by a hospital's distance from the benchmark in the baseline period: $p_i^m = \text{Round}\left[10\dfrac{y_i^m - y_{i0}^m}{m_b - y_{i0}^m} - 0.5\right]$, where y_{i0}^m is the score for measure m for hospital i in the baseline period.

Additionally, *consistency points* c_i for patient experience of care domain are calculated as the lowest of the 8 dimension scores d_i^m:

$c_i = \text{Round}\left[20 \min_m\{d_i^m\} - 0.5\right]$, where $d_i^m = \dfrac{y_i^m - m_f}{m_a - m_f}$, m_f is the floor for measure m and $m = 1, ..., 8$.

The score for clinical process of care domain is the sum of the values for its twelve quality measures, divided by the total potential score (12·10) and translated into percentage points:

$d_{i1} = \dfrac{\sum_{m=1}^{12} \max\{a_i^m, p_i^m\}}{120} \cdot 100.$

In case of eight measures of patient experience of care, the domain score is the sum of the values for each

[4] Setting the actual value of s slightly above 2 may be explained by an increase in the number of data-reporting hospitals, whose *tps* may not be present for the baseline period but on average is expected to be higher than the historic mean.

[5] Additionally, outcome of care domain is added for FY 2014 and efficiency domain for FY 2015.

measure plus consistency points, divided by total potential score for quality measures (8·10) plus maximum value for consistency points (20) and translated into percentage:[6]

$$d_{i2} = c_i + \sum_{m=1}^{8} \max\{a_i^m, p_i^m\}$$

Finally, the total performance score of each hospital is a weighted sum of domain scores:
$tps_i = 0.7 \cdot d_{i1} + 0.3 \cdot d_{i2}$.

It should be noted that the formula for improvement points is targeted exclusively at hospitals *below* the benchmark. Hospitals already performing above the benchmark obtain the maximum of achievement points, and hence, get the maximum potential value for achievement or improvement points. Consequently, the value-based purchasing reward schedule does not provide any financial incentives for further improvement of the best-performing hospitals. Moreover, value-based reimbursement may be regarded as a stepwise function, since the benchmark value is established at the mean of the top decile. So a group of hospitals above the benchmark is guaranteed the maximum score on corresponding measures.

Eligibility criteria for participation in value-based purchasing are: 1) at least 100 surveys in the patient experience of health care; 2) data on at least 4 measures of clinical process of care, with at least 10 respondents on each measure; 3) acute care hospital; 4) outside of Puerto Rico.

Total performance score was calculated based on FY2012 survey, and hospital adjustment coefficient was established for FY2013. Out of eligible 96% submitted data in the prereform period (FY2012), 89% joined in the first year (FY2013), 92% joined in the second year (FY2014). Compliers with the reform are larger hospitals in terms of number of beds or hospital budget. We do not observe any "scale" economy in terms of quality performance and hospital size. Moreover, compliers are hospitals with relatively *lower* scores for most quality measures, and higher standard deviation and lower minimum values for patient experience of care measures. But generally the differences between compliers and non-compliers are statistically insignificant. Therefore, the main reason for non-participation is not related to the performance, but rather the monetary equivalent of 1–2% of Medicare's hospital budget.

The *nationwide* quality-performance reimbursement started in the U.S. with the Hospital Quality Incentive Demonstration (HQID), when 33 quality measures for five clinical conditions (heart failure, acute myocardial infarction, community-acquired pneumonia, coronary-artery bypass grafting, an hip and knee replacement) were accumulated from voluntarily participating hospitals.[7] 266 out of these quality-reporting 613 hospitals opted for the pay-for-performance project (initially established for 2003-2006, and later extended to 2007-2009). The project provided respectively 2% and 1% bonus payments for hospitals in the top and second top deciles of each quality measure. On the other hand, hospitals in the bottom two deciles (as of the end of the third year of the project) were to receive 1-2% penalties. It should be noted that HQID redistributed funds between top and bottom hospitals, while value-based purchasing applies deductions or rewards to all hospitals. Therefore, the potential impact of value-based purchasing might be expected to be higher than that of HQID (Kahn et al. (2006)).

Overall, the financial incentives helped improve the quality of the participant hospitals, but the improvement was inversely related to baseline performance (Lindenauer et al. (2007)). Moreover, low-quality hospitals required most investment in quality increase, yet, they were not financially stimulated (Rosenthal et al. (2004)). This outcome might have been the reason for an extension of reimbursement rules within value-based purchasing into achievement and improvement points.

The accumulation of the measures within the Hospital Quality Incentive was followed by the launch of the Surgical Care Improvement Project (SCIP) and Hospital Consumer Assessment of Healthcare Providers

[6] So 100 in denominator and nominator cancel out.
[7] The Centers for Medicare and Medicaid Services (CMS) - Premier database.

(HCAHPS). HCAHPS was the first national standardized survey with public reporting on various dimensions of patient experience of care (HCAHPS online (2013)), and its measures are the basis for the patient experience of care domain in value-based purchasing. The measures of the clinical process of care domain are collected within Hospital Inpatient Quality Reporting (IQR) program. These are measures for acute clinical conditions stemming from the Hospital Quality Incentive (i.e. AMI, heart failure, pneumonia), as well as measures from the Surgical Care Improvement Project and Healthcare Associated Infections.

3.2 Length-of-stay performance in Japan

An inpatient PPS as a means of curtailing explosive costs of the health care system was piloted in Japan in 1990 (See a review of the Japanese health care system in Appendix A). Inclusive per diem rates (unadjusted for case-mix) were employed in 50% of geriatric hospitals, which satisfied the required staffing criteria (MHLW 2012a, Ikegami (2005), Okamura et al. (2005)). Later a system with per case payments was tested at 10 acute care hospitals in 1998-2004 (Kondo and Kawabuchi (2012)). However, owing to high diversity of medical treatment patterns, the effect of this full PPS was ambiguous and the system was not expanded nationwide (Kondo and Kawabuchi (2012), Okamura et al. (2005)). Therefore, a modified *per diem* payment system was approved for nationwide use. The per diem rates were originally set on the basis of 1860 homogeneous diagnosis groups, which covered about 90% of admissions at 82 forerunner hospitals[8] in 2003 (Ikegami (2005)). Subsequently, the number of diagnosis groups was adjusted and the share of homogeneous diagnosis groups steadily rose. There was an increase in the number of Japanese hospitals, joining the PPS voluntarily in 2004-2014.[9] As of April 2014, there were 2873 diagnosis groups and 2309 diagnosis-procedure combinations, and 21.1% of acute care hospitals, accounting for 54.7% of acute care hospital beds in Japan, are financed using length-of-stay performance-based reimbursement (Ministry of Health, Labor and Welfare (2014b)).

Similarly to the U.S. value-based purchasing, the Japanese performance-based remuneration guarantees that hospitals performing better than the mean national level remain budget neutral. The amount of the daily inclusive payment for each diagnosis-procedure combination (DPC) is flat over each of the three consecutive periods: *period I* represents the 25-percentile of length of stay calculated for all hospitals submitting data to MHLW; *period II* contains percentiles from 25 to mean length of stay; and *period III* includes two standard deviations from the mean (Ministry of Health, Labor and Welfare (2014b)).[10] Per diem payment P^I in *period I* is set higher than the average per diem payment $\bar{P}$ in the pre-reform schedule, so that the gain in producer surplus in *period I* equaled the loss in producer surplus in *period II* (areas A and B on Figure 1 are equal).[11] The per diem reimbursement in *period III* is 10-15% lower than in *period II*, so hospitals with value of length of stay worse than the national mean suffer financial losses. Overall, the schedule creates incentives for shorter lengths of stay, establishing financial rewards in *period I*.

The 14-digit code for each DPC incorporates diagnosis, medical algorithm, procedure, and co-morbidity. The first 6 digits account for diagnosis: 2 digits for major diagnostic category, and 4 digits for the name of disease (Appendix B). Digit 7, which indicated for the type of hospitalization,[12] is not in use since 2006. Digit 8 (age, birth weight, Japan coma scale, burn index) is employed for selected MDCs.[13] Digits 9 and 10 indicate the type of operation. Digits 11 and 12 code additional surgical procedures and adjuvanct

[8] 80 university hospitals and two national centers, providing high-technology health care.

[9] Owing to stepdown per diem schedule, the new system is particularly attractive to hospitals with length of stay less than mean nationwide length of stay.

[10] After the end of *period III*, hospitals are reimbursed according to the fee-for-service system.

[11] So payment in *period I* is 15% higher than p for a standard DPC, 10% higher for a DPC with low medical cost at the beginning of the treatment, and varies for a DPC with high medical cost at the beginning of the treatment.

[12] "1" – examination and tests, "2" – study (educational).

[13] "Age" is used for MDCs 04,06,13,14,15,18; "birth weight" applies to MDC14; "Japan coma scale" – to MDC01; "Burn index" – to MDC16.

therapy, respectively. Digits 13 and 14 contain sub-codes for co-morbidities and severity (Figure C.1). Diagnoses are coded according to ICD-10 (with minor aggregation or disaggregation of diagnoses within ICD's Major Diagnostic Categories, MDCs, Table C.1) and procedures are classified on the basis of the Japanese Procedure Code, commonly used under fee-for-service reimbursement (Ministry of Health, Labor and Welfare (2014a), Matsuda et al. (2008)).

Owing to increase in standardization of medical treatment patterns, the 25th percentile of length of stay is gradually approaching mean value of length of stay in most diagnosis groups. So MHLW introduced a change in the pricing schedule: starting fiscal year 2012 the rate in *period I* cannot be applied to the number of days exceeding 50% of the mean length of stay (Ministry of Health, Labor and Welfare (2012a)). In economic terms this relates to a change in benchmarking, since the length of *period I* is essentially reestablished as $\min\{25th \text{ percentile}, 0.5\text{mean}\}$. Moreover, the length of *period I* is decreased to only one day for 22 DPCs with particularly high medical cost at the beginning of the treatment (Ministry of Health, Labor and Welfare (2012b)). The policy change may be viewed as an attempt to diminish the adverse effects of degressive tariff-setting and move towards best-practice rate setting.

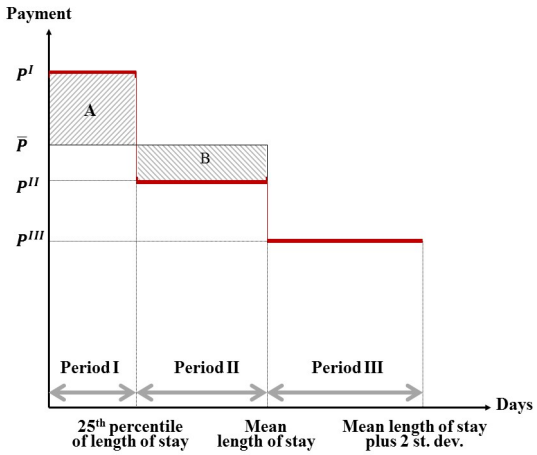

Figure 1: **Stepdown schedule for a standard DPC before 2012**
Source: Ministry of Health, Labor and Welfare (2014b)

The Japanese version of inpatient PPS is a mixed system. The two-part tariff is the sum of a diagnosis-procedure combination and fee-for-service components, with approximate shares are 0.7 and 0.3 respectively (Okamura et al. (2005)). [14] The two-component system may be justified in part by the historically developed variety of practice patterns in Japanese hospitals (Hamada et al. (2012), Campbell and Ikegami (1998)). The Japanese two-part tariff resembles the German PPS in 1996-2003, where the per diem fee was a sum of a department-specific prospective component for medical costs and a hospital-specific retrospective component for nonmedical costs (Busse and Schwartz (1997)).

[14] The DPC component reimburses the cost of hospital fee, examinations, diagnostic images, pharmaceuticals, injections, and procedures worth less than 10,000 yen. The fee-for-service component covers medical teaching, surgical procedures, anaesthesia, endoscopies, radioactive treatment, pharmaceuticals and materials used in operating theatres, as well as procedures costing more than 10,000 yen (Ministry of Health, Labor and Welfare (2014b), Yasunaga et al. (2005)).

4 Methodology

4.1 Intuition

A hospital is regarded as a profit-maximizing supplier of health care in a certain volume and to a certain level of quality, and a hospital's objective function includes a benefit of providing health care to patients (Ma (1998), Ellis and McGuire (1996), Hodgkin and McGuire (1994), Laffont and Tirole (1993), Ellis and McGuire (1986)).

4.1.1 Fee-for-service reimbursement

The behavior of a hospital under a *a fee-for-service* reimbursement for treating a patient with a given disease can be analyzed with Grabowski et al. (2011) approach, defining a hospital's profit as $\pi = (p(e) - c(e)) \cdot LN(B(L,e))$, where per diem payment $p(e)$ depends on intensity e (which may be also regarded as hospital's efforts), $c(e)$ is cost[15] ($c_e > 0, c_{ee} > 0$), $p(e) - c(e)$ is the margin of the service, $L \cdot N(B(L,e))$ is health care volume, L is length of stay, benefit to patient B is a function of L and e ($B_e > 0$), $N(B)$ is the number of discharges ($N_B > 0$). The first order conditions derived in Grabowski et al. (2011) are as follows:

$$\partial\pi/\partial L = 0 \implies N + LN_B B_L = 0, \tag{2}$$

and since $N > 0$ and $N_B > 0$, it must hold that $B_L < 0$.

$$\partial\pi/\partial e = 0 \implies (p_e - c_e)N + (p - c)N_B B_e = 0. \tag{3}$$

4.1.2 Quality-performance

In case of the U.S. inpatient prospective payment system the price p is paid for the whole episode of treatment. Under *quality-based* reimbursement p becomes a function of quality: $p = p(q), p_q \geq 0$. Quality depends on hospital's effort: $q = q(e), q_e \geq 0$. Benefit to patient may be viewed as a function of quality $B = B(q)$, where it should hold that $B_q \geq 0, B_{qq} \leq 0$. The number of discharges becomes a function of $B(q)$: $N = N(B(q)), N_B \geq 0$.

So the profit function modifies to $\pi = \big(p(q(e)) - c(e)\big)N(B(q(e))$. The first order condition with respect to e is:

$$\frac{\partial\pi}{\partial e} = (p_q q_e - c_e)N + (p - c)N_B B_q q_e = 0 \tag{4}$$

Case 1. $p_q = 0$, which corresponds to the best-performing hospitals. If N is bounded from above and $N_q = N_B B_q$ is small (owing to inability to accept more patients given the number of beds and bed occupancy rate), then $\dfrac{\partial\pi}{\partial e} \leq 0$ and we have a corner solution with $q_e = 0$.[16] Consequently, best-performing hospitals may have no incentive to improve q.

Intuitively, if a hospital provides an optimal treatment level, the marginal benefit of additional services may be zero for the consumer, so no improvement may be expected (Moreno-Serra and Wagstaff (2010)). Moreover, the consumer demand for quality may be inelastic (Miraldo et al. (2011)), resulting in $N_q = 0$ and lower optimal price. So best-performing hospitals may not have incentives to increase their quality.

Case 2. $p_q \geq 0$, which reflects regulator's payment rules for the rest of the hospitals. Then equation (4) has an additional positive term $p_q q_e N$ and it is possible that $q > 0$.

[15] Here and in the model for the Japanese PPS $c(\cdot)$ denotes per diem cost, while $c(\cdot)$ is the cost for the whole episode of treatment in case of the U.S. PPS.

[16] Or whatever effort that keeps these hospitals in the group with $p_q = 0$.

4.1.3 Length-of-stay performance

To model a Japanese per diem payment system with *length-of-stay* performance, we regard p as a function of both intensity and length of stay: $p = p(e, L)$. The per diem payment in Japan system decreases with length of stay, so $p_L < 0$. The profit function becomes $\pi = (p(e, L) - c(e)) \cdot LN(B(L, e))$ so (3) does not change, but (2) modifies to:

$$\frac{\partial \pi}{\partial L} = p_L + (p - c)(1/L + N_B B_L/N) = 0 \tag{5}$$

Equation 5 gives $B_L = -N\dfrac{p_L L + (p - c)}{(p - c)N_B L}$. Since $N > 0$, $N_B > 0$, $L > 0$, $signB_L = -sign\left(\dfrac{p_L L}{(p - c)} + 1\right)$.

Case 1. Hospitals with the shortest length of stay are given $p > c$ under the Japanese payment schedule. Owing to $p_L < 0$, these hospitals may have $p_L L/(p - c) < -1$ and consequently, sign $B_L > 0$. So Japanese hospitals with the shortest L may increase their length of stay.

Case 2. If L is large, the absolute value of p_L may be very small. So $p_L L/(p - c)$ becomes less than unity, leading to $B_L < 0$. Consequently, hospitals with longer L have incentives to shorten length of stay.

4.2 Comparative statics

4.2.1 Quality performance

Profit-maximizing model

Assume i-th hospital chooses quality $q \in [\underline{q}, \bar{q}] \subset R_+$ to maximize its profit π_i expressed as

$$\pi_i = (d - \alpha t + \gamma t q - c(q, i)) \cdot N(q).$$

Here $N(q) > 0$ is the mass of clients served (increasing in quality), $d > 0$ is a standard prospective payment per patient with a given diagnosis, $c(q, i)$ is the individualized (increasing in q and i) cost function of this hospital, $t \geq 0$ is some governmental stimulating treatment, $\alpha > 0$, $\gamma > 0$. Whenever the government chooses zero treatment, the hospital gets only standard payment, while the bigger is scale t of treatment, the stronger is the incentives regulation. The coefficient α de-stimulates hospital's activity per se, while coefficient γ stimulates quality. In view of this contradiction, what will be their joint impact on quality? Would firms with unequal costs respond differently?

Naturally, the answer depends upon (increasing differentiable) demand-for-quality function $N(q)$ and (increasing differentiable) cost function $c_i(q)$. If N were a constant, parameter α would have no impact and only stimulation γ would play a positive role, as well as in the case of positive parameter α. Depending upon curvature of demand N and cost c_i, the profit function π_i can be concave or not, anyway having some argmaxima on compact domain $[\underline{q}, \bar{q}]$. Three cases may arise: left corner solution $q^* = \underline{q}$, inner solution $q^* \in (\underline{q}, \bar{q})$ and right corner solution $q^* = \bar{q}$. For inner solution, profit π_i must be locally concave at least at this point, otherwise concavity is not needed for analysis. We need only FOC determining all local minima and maxima:

$$\pi'_{i(t)}(q) = (\gamma t - c'(q, i)) \cdot N(q) + (d - \alpha t + \gamma t q - c(q, i)) \cdot N'(q) \lessgtr 0$$

Whenever this expression is negative (positive) everywhere on $(\underline{q}, \bar{q})$, it generates the left (right) corner solution. But when the expression changes the sign, an inner solution may occur (see the above computations of the maximum). A certain treatment level $\hat{t} > 0$ has an impact on firm i only in the third case, i.e., when the left (or right) corner solution does not persist on interval $t \in [0, \hat{t}]$, i.e., curve $\pi'_{i(t)}(q)$ does change its sign at least somewhere on $(\underline{q}, \bar{q})$.

The unconstrained local argmaxima of this (monotone or non-monotone) curve are such that its intersections q^* with zero are where FOC curve $\pi'_{i(t)}(q)$ crosses zero *downwards*. When the highest of such q^*, the *global* unconstrained argmaximum lies to the left (right) from interval $(\underline{q}, \bar{q})$, then the left (right) constrained corner solution occurs. Concerning *local* argmaximum at the inner solution, note that

If the cost function c is supermodular (stronger firms have higher number), then stronger firms produce higher quality: $c'_i > c'_j \Rightarrow q_i > q_j$. Proof: see Milgrom and Shannon (1994) and Milgrom and Roberts (1995).

In other words, any unconstrained local argmaximum q^* shifts to the right (left) whenever the treatment parameter t shifts the whole FOC curve $\pi'_{i(t)}(q)$ upward (downward), which means supermodularity (complementarity) between variables (q, t) in the profit function π.

Looking at data in the below empirical section, we may conjecture that for weak firms (high marginal cost c'_i of generating quality) supermodularity does take place. Indeed, weak firms increase their effort in response to governmental treatment, whereas the opposite effect prevails for strong firms.

Alternatively, to infer comparative statics in response to treatment t, we algebraically find a relation between t and q in the profit function through differentiating the FOC in t:

$$\frac{\partial^2 \pi_{i(t)}(q)}{\partial t \partial q} = \gamma \cdot N(q) + (-\alpha + \gamma q) \cdot N'(q) \lessgtr 0$$

which (under $N'(q) > 0$) entails condition for positive impact of t on q as

$$n(q_i, \frac{\alpha}{\gamma}) \equiv \frac{N(q_i)}{N'(q_i)} + q_i - \frac{\alpha}{\gamma} > 0,$$

whereas negative impact occurs under opposite inequality. Both sides are positive and the relation says that under sufficiently low fraction $\frac{\alpha}{\gamma}$ (low de-stimulation α and high stimulation γ) all firms, independently of individualized quality q_i, would respond positively to treatment.

If $n(q_i, \frac{\alpha}{\gamma}$ changes its sign at $(\underline{q}, \bar{q})$, when some firms demonstrate positive function $n(q_i^{*0}, 0) > 0$ in the no-treatment state of the world ($t = 0$) but some have negative value $n(q_j^{*0}, 0) > 0$. Then it is possible (and guaranteed under small treatment t) that new situation $t > 0$ generates different response to treatment, i.e., $q_i^{*t} > q_i^{*0}$, $q_j^{*t} < q_j^{*0}$.

In particular, converging behavior (weak firms increase but strong firms decrease their quality) requires that $\frac{N(q_i)}{N'(q_i)} + q_i$ is a decreasing function on at least some upper sub-interval of $(\underline{q}, \bar{q})$. The required condition is

$$\frac{d\left(\frac{N(q)}{N'(q)} + q\right)}{dq} = 1 - \frac{N(q)N''(q)}{N'^2(q)} + 1 < 0$$

Weak first derivative and strong positive second one, i.e. fraction of second and first elasticities bigger than 2:

$$\frac{qN''(q)}{N'(q)} \cdot \frac{N(q)}{qN'(q)} > 2.$$

Note, however, the profit-maximizing model induces very specific assumptions about the form of the demand function for explaining a converging case.

At the same time, divergence of behavior is observed when $n'(\cdot)$ is an increasing function, because parameter α de-stimulates all firms equally, while parameter γ stimulates strong ones (those with high quality) stronger than weak ones, so, they should increase quality. For instance, when $N(q) = q^2$, we get an increasing function $\frac{N(q_i)}{N'(q_i)} + q_i = \frac{q_i^2}{2q_i} + q_i = (0.5q_i + 1)q_i$.

Consequently, the below section proposes an alternative modeling approach, which provides for converging

behavior.

Utility-maximizing model

Empirical comparison of hospitals' behavior before and after introduction of quality stimulation has shown somewhat paradoxical fact. Though the reform introduced a reward to a hospital, positively dependent upon some measure of quality, that replaced zero dependence but the response was partially negative.[17] The quality decreased at least for high-quality group of hospitals. If we perceive hospitals as risk-neutral profit-maximizing agents, rationalization of such negative response is hardly possible.

By contrast, rationalization becomes possible if we think of hospitals as risk-averse agents or just agents with decreasing marginal utility of money. To motivate such approach, we can say that quality is not chosen directly by top officials but rather rather by a large group of personnel. Personnel compares the probability of being fired (in the case of bankrupt hospital) with personal effort to maintain quality. In this respect, the reform introduced two novelties: (1) in essence, decreased the level of reward for qualitatively weak hospitals and increased it for strong ones; (2) introduced a positive dependence of reward for everybody. We would argue, that these two forces can struggle with each other and the first can outweigh the second one for strong agents: they become de-stimulated rather than stimulated.[18]

Under a principal-agent approach, an agent has an increasing strictly concave valuation function for monetary reward $v(r)$ and bears own effort measured by quantity or quality $q > 0$ to get reward, with type-specific coefficient θ:[19] $U = v(r) - q/\theta$. Principal does observe quality (imperfect observation would slightly complicate the model but does not change the outcome), and suggests a linear contract in the form

$$r(q) = bq + t\alpha + t\gamma q.$$

Here bq is some initial (pre-reform) reward for quality (b may be viewed as a derivative of the demand $N'_q(q)$ under linear $N(q)$), $t \in \{0, 1\}$ is the indicator of treatment or not (reform or not). Positive or negative α is the treatment reward for the fact of existence and $\gamma \geq 0$ is the treatment additional stimulation coefficient. Let us show that these two principal's instruments have the opposite impact on behavior when being positive. The coefficient stimulates but fixed payment de-stimulates. Indeed, when satisfaction v is measured in effort (which is identical to quality in our terms) taken with type-specific coefficient θ, then agent solves the problem

$$\max_q U \equiv v(bq + t\alpha + t\gamma q) - q/\theta$$

and its FOC is

$$\frac{\partial U}{\partial q} = v'(bq + t\alpha + t\gamma q)\,(b + t\gamma) - 1/\theta = 0,$$

the higher type the higher quality because v' is a decreasing function.

$$q_\theta : \; v'(bq + t\alpha + t\gamma q_\theta)\,(b + t\gamma) = 1/\theta.$$

Sub-modularity between (α, q) is shown as $\frac{\partial^2 U}{\partial q \partial \alpha} = v''(\cdot)t < 0$, which is true for concave v everywhere. Sub-modularity implies that α can have de-stimulating effect.

For low-type hospitals their type-specific chosen quality q_θ was lower than for high types. Then treatment

[17] See below section with quantile regressions and clinical process of case measures.

[18] Likewise, in response to doubling wage per hour on African copper mines, the labor supply decreased exactly twice, because of satiation with $10 a day.

[19] The parameter is similar to hospital's index i in the previous section. It does not need to be discreet, but can be continuous on some interval

$(t = 1)$ with negative α replaced their previous reward bq with smaller magnitude $bq+\alpha+\gamma q_\theta < bq$ and added stimulation coefficient γ, thereby twice stimulating them. By contrast, the high-type (high-performing, i.e. high-quality) hospitals become richer (de-stimulation effect) but get added stimulation coefficient γ, which is insufficient to outweigh the income effect when

$$\frac{\partial^2 U}{\partial q \partial t} = \frac{\partial}{\partial t}[v'(bq_\theta + t\alpha + t\gamma q_\theta)\gamma] = v''(\cdot)[\alpha + \gamma q_\theta] < 0,$$

i.e., when treatment rule $(\alpha+\gamma q_\theta)$ applied to initial behavior q_θ brings surplus rather than loss to a hospital.

This argument shows that utility-maximizing hospitals can show convergence of behavior in response to treatment: low-type increase their effort but high-type decrease

4.2.2 Length-of-stay performance

Profit-maximizing model

Assume i-th hospital chooses length of stay $L \in [\underline{L}, \bar{L}] \subset R_+$ to maximize its profit π_i expressed as

$$\pi_i = (d + \alpha t L - \gamma t L - c(L, i))\, L \cdot N(L).$$

The first order condition is

$$\pi'_L = \big(-\gamma t - c'(L, i)\big) L N(L) + \big(d + \alpha t - \gamma t L - c(L, i)\big)\big[N(L) + L N'_L(L)\big]$$

The impact of treatment t may be inferred through a mixed partial derivative of π in case of a corner solution (supermodularity or implicit function theorem in a smooth continuous case):

$$\frac{\partial^2 \pi_{i(t)}(L)}{\partial t \partial L} = -\gamma L N(L) + (\alpha - \gamma L)\,[N(L) + L N'_L(L)]$$

Here $\alpha, \gamma > 0$, $N, L > 0$, so we may divide by $\gamma L N$ and write a condition for a negative impact of of t on L as

$$n(L_i, \frac{\alpha}{\gamma}) \equiv -2 + \frac{\alpha}{\gamma L} + \frac{\alpha N'(L_i)}{\gamma N(L_i)} - \frac{L N'(L_i)}{N(L_i)} < 0,$$

In particular, converging behavior (weak firms decrease length of stay, while strong firms increase it) requires:

$$\frac{dn}{dL} < 0 \iff -\frac{\alpha}{\gamma L^2} < \frac{N'(L)}{N(L)},$$

which is plausible, since $N'(L)$ is likely to be negative at least at some subinterval of $[\underline{L}, \bar{L}]$.

Utility-maximizing model

An agent has an increasing strictly concave valuation function for monetary reward $v(r)$ and bears own effort measured by quantity $L > 0$ to get reward, with type-specific coefficient θ : $U = v(r) - \theta L$. Principal suggests a linear contract in the form

$$r(L) = bL + t\alpha - \gamma t L.$$

Here bL is some initial (pre-reform) payment (b may be viewed as a derivative of the demand $N'_L(L)$), $t \in \{0, 1\}$ is the indicator of treatment or not (reform or not).

The FOC is

$$\frac{\partial U}{\partial L} = v'(bL + t\alpha - \gamma tL)\,(b - \gamma t) - \theta = 0,$$

which determines the type-specific length of stay

$$L_\theta :\ v'(bq + t\alpha - \gamma t L_\theta)\,(L_\theta - \gamma t) = \theta.$$

For low-type hospitals their type-specific L_θ was higher than for high types (longer length of stay at the worst hospitals). Then the reform ($t = 1$) replaced their previous reward bL with smaller magnitude $bL + \alpha - \gamma L_\theta < bL$ and added stimulation coefficient γ, thereby twice stimulating them. High-type (high-performing, i.e. low length of stay) hospitals become richer (destimulation effect) but get stimulation coefficient γ, which is sufficient to outweigh the income effect, since

$$\frac{\partial^2 U}{\partial L \partial t} = \frac{\partial}{\partial t}[v'(bL_\theta + t\alpha - \gamma t L_\theta)\gamma] = -v'(\cdot)\gamma + v''(\cdot)[\alpha - \gamma L_\theta] > 0$$

So utility-maximizing model with some satiation does not work in case of Japanese payment scheme.

4.3 Empirical approach

Our analysis assumes that a hospital strongly adheres to its practice patterns (Campbell and Ikegami (1998)). So for each group of diagnoses the value of performance measures depends on the value of the variable in the previous period. We model performance-based rate-setting using quantile regressions, which allows incorporating percentile-dependent price schedule and heterogeneity. As robustness check, we compare our findings to the estimates with conventional least squares approach.

4.3.1 Dynamic quantile regressions

Random effects model

The model is a longitudinal version of Chernozhukov and Hansen (2008), specified as:

$$y_{it} = \mathbf{d}'_{it}\boldsymbol{\alpha}(u_{it}) + \mathbf{x}'_{it}\boldsymbol{\beta}(u_{it}) \tag{6}$$

$$\mathbf{d}'_{it} = \delta(\mathbf{x}_{it}, \mathbf{z}_{it}, \nu_{it}) \tag{7}$$

$$\tau \mapsto \mathbf{d}'_{it}\boldsymbol{\alpha}(\tau) + \mathbf{x}'_{it}\boldsymbol{\beta}(\tau) \tag{8}$$

where τ denotes the value of a given quantile for conditional distribution of the dependent variable y for observation i at period t, $\mathbf{d}$ is a vector of endogenous variables, $\mathbf{x}$ is a vector of exogenous variables (in our case, hospital characteristics and annual dummies), $\mathbf{z}$ is a vector of instruments, ν_{it} is statistically dependent on u_{it}, $u_{it} \perp (\mathbf{x}_{it}, \mathbf{z}_{it}) \sim U[0,1]$, $i = 1, ..., N$, $t = 1, ..., T$. Here y_{it} is hospital-level quality measure in the analysis with the Medicare's data and average length of stay for each major diagnostic category or diagnosis-procedure combination in case of the Japanese data.

A consistent estimation procedure (Galvao (2011), Chernozhukov and Hansen (2008)) involves minimizing the weighted quantile regression objective function

$$Q_{NT}(\tau, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) := \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} \rho_\tau(y_{it} - \mathbf{d}'_{it}\boldsymbol{\alpha} - \mathbf{x}'_{it}\boldsymbol{\beta} - \boldsymbol{\phi}'_{it}\boldsymbol{\gamma})v_{it} \tag{9}$$

where $\rho_\tau(u) = u(\tau - I(u \le 0))$ is the least absolute deviation loss function (Koenker and Bassett (1978)),

$\boldsymbol{\phi}_{it} = f(\mathbf{x}_{it}, \mathbf{z}_{it})$ and $v_{it} := v(\mathbf{x}_{it}, \mathbf{z}_{it})$ are weights.

Using the inverse quantile regression approach, the first step requires obtaining

$$\big(\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}, \tau), \hat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \tau)\big) := \underset{\boldsymbol{\beta}, \boldsymbol{\gamma}}{\operatorname{argmin}} Q_{NT}(\tau, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \tag{10}$$

Second, the value of $\boldsymbol{\alpha}$ that minimizes $\hat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \tau)$ is found as (Chernozhukov and Hansen (2004), eq.3.2):

$$\hat{\boldsymbol{\alpha}}(\tau) = \underset{\boldsymbol{\alpha} \in \mathcal{A}}{\operatorname{argmin}} W(\boldsymbol{\alpha}), W(\boldsymbol{\alpha}) := \hat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \tau)' \hat{A}(\boldsymbol{\alpha}) \hat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \tau) \tag{11}$$

where $A(\boldsymbol{\alpha})$ is uniformly positive definite matrix in compact parameter set $\mathcal{A}$.

The variance-covariance matrix $\mathbf{J}(\tau)^{-1}\mathbf{S}(\tau, \tau')[\mathbf{J}(\tau)^{-1}]'$ of $\hat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \tau)$ is estimated as (Chernozhukov and Hansen (2006), eq.3.11-3.14):

$$\hat{\mathbf{S}}(\tau, \tau') = (\min(\tau, \tau') - \tau\tau') \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \hat{\boldsymbol{\psi}}'_{it}(\tau) \hat{\boldsymbol{\psi}}_{it}(\tau') \tag{12}$$

$$\hat{\mathbf{J}}(\tau) = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} I(|\hat{\epsilon}_{it}(\tau)| \leq h_{NT}) \hat{\boldsymbol{\psi}}'_{it}(\tau) [\mathbf{d}'_{it}, \mathbf{x}'_{it}] \tag{13}$$

where $\hat{\epsilon}_{it}(\tau) \equiv y_{it} - \mathbf{d}'_{it}\hat{\boldsymbol{\alpha}}(\tau) - \mathbf{x}'_{it}\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\phi}'_{it}\hat{\boldsymbol{\gamma}}(\tau)$, $\boldsymbol{\psi}_{it}(\tau) \equiv v_{it}(\tau) \cdot [\boldsymbol{\phi}'_{it}(\tau), \mathbf{x}'_{it}]$, and bandwidth h_{NT} is chosen so that $h_{NT} \to 0$ and $NTh_{NT}^2 \to \infty$.

We modify $\hat{\mathbf{S}}$ adding the Parente and Santos Silva (2015) approach for clustered standard errors in quantile regression, regarding each observation as a longitudinal cluster. The scores of the objective function $s_{it}(\tau)$ are computed as a piecewise derivative:

$$s_{it}(\tau) = \frac{\partial \rho_\tau(\epsilon_{it}(\tau))}{\partial [\boldsymbol{\gamma}', \boldsymbol{\beta}']} = -[\boldsymbol{\phi}'_{it}(\tau), \mathbf{x}'_{it}] \chi_\tau(\epsilon_{it}(\tau)), \tag{14}$$

where $\chi_\tau(\epsilon_{it}(\tau)) = \tau - I(\epsilon_{it}(\tau) < 0)$. The scores have the zero mean at the true value of parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (where $\boldsymbol{\gamma} = 0$), conditional on $\boldsymbol{\phi}_{it}$ and $\mathbf{x}_{it}$. This is equivalent to conditioning on $\mathbf{z}_{it}$ and $\mathbf{x}_{it}$, since $\boldsymbol{\phi}$ is a function of $\mathbf{x}$ and $\mathbf{z}$. Indeed, the direct computation of the mean gives:

$$\begin{aligned} E\big(s_{it}(\tau) \mid \mathbf{x}_{it}, \mathbf{z}_{it}\big) &= E\big(-[\boldsymbol{\phi}'_{it}(\tau), \mathbf{x}'_{it}]\theta_\tau(\epsilon_{it}(\tau))\big) \propto \tau E\big(I(\epsilon_{it}(\tau) \geq 0) \mid \mathbf{x}_{it}, \mathbf{z}_{it}\big) - (1-\tau) E\big(I(\epsilon_{it}(\tau) < 0) \mid \mathbf{x}_{it}, \mathbf{z}_{it}\big) \\ &= \tau Pr\{y_{it} - \mathbf{d}'_{it}\boldsymbol{\alpha} - \mathbf{x}'_{it}\boldsymbol{\beta} \geq 0 \mid \mathbf{x}_{it}, \mathbf{z}_{it}\} - (1-\tau) Pr\{y_{it} - \mathbf{d}'_{it}\boldsymbol{\alpha} - \mathbf{x}'_{it}\boldsymbol{\beta} < 0 \mid \mathbf{x}_{it}, \mathbf{z}_{it}\} \\ &= \tau(1-\tau) - (1-\tau)\tau = 0, \end{aligned}$$

since $Pr\{y_{it} - \mathbf{d}'_{it}\boldsymbol{\alpha} - \mathbf{x}'_{it}\boldsymbol{\beta} \geq 0 \mid \mathbf{x}_{it}, \mathbf{z}_{it}\} = 1 - \tau$. So according to the assumption 2 in Parente and Santos Silva (2015) equation (12) modifies to:

$$\mathbf{S}(\tau) = E\Big[\sum_{s=1}^{T} \sum_{t=1}^{T} s_{is}(\tau) s_{it}(\tau)'\Big] \tag{15}$$

and $\mathbf{J}(\tau) = \sum_{t=1}^{T} E[\hat{\boldsymbol{\psi}}_{it}(\tau) \boldsymbol{\phi}'_{it} f(0|\mathbf{x}_{it}, \boldsymbol{\phi}_{it})]$, where $f(\epsilon|\mathbf{x}_{it}, \boldsymbol{\phi}_{it})$ is the density of the conditional distribution of $F(\epsilon_{it}(\tau)|\mathbf{x}_{it}, \boldsymbol{\phi}_{it})$.

As is shown in the Parente and Santos Silva (2015), the consistent estimators become:

$\hat{\mathbf{S}}(\tau) = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\boldsymbol{\psi}_{it}\boldsymbol{\psi}'_{is}\chi_\tau(\hat{\epsilon}_{it}(\tau))\theta_\tau(\hat{\epsilon}_{is}(\tau))$ and $\hat{\mathbf{J}}(\tau) = \frac{1}{2\hat{c}_N N}\sum_{t=1}^{T}\sum_{i=1}^{N} I(|\hat{\epsilon}_{it}(\tau)| \leq \hat{c}_N)\hat{\boldsymbol{\psi}}_{it}(\tau)[\mathbf{d}'_{it}, \mathbf{x}'_{it}]$, where $\hat{c}_N = \kappa[\Phi^{-1}(\tau + h_{NT}) - \Phi^{-1}(\tau - h_{NT})]$, with κ equal to median absolute deviation of the τ-th quantile regression residuals and h_{NT} defined in Koenker and Machado (1999).

Note that Wooldridge (2007) proposes similar use of scores for correction of $\hat{\mathbf{S}}$ in time-series estimates and Wang and He (2007) derive asymptotic properties of rank scores tests in a multiple quantile model.

"Locational shift" fixed effects model

Denote $\tilde{y}_{it} = y_{it} + \eta_i$, $\tilde{\tilde{\mathbf{x}}}_{it} = [\mathbf{d}_{it}, \mathbf{x}_{it}]$. Canay (2011) showed the consistency of a two-step estimator for the below system with exogenous $\tilde{\tilde{\mathbf{x}}}_{it}$:

$$y_{it} = \tilde{\tilde{\mathbf{x}}}'_{it}\boldsymbol{\theta}(u_{it}) + \eta_i \tag{16}$$

$$\tau \mapsto \tilde{\tilde{\mathbf{x}}}'_{it}\boldsymbol{\theta}(\tau) \tag{17}$$

under η_i independent of u_{it} (assumption 1) and $u_{it} \perp (\tilde{\tilde{\mathbf{x}}}_{it}, \eta_i)$ (assumption 2). At the first stage, a $\sqrt{NT}$ least squares consistent estimator of $\hat{\boldsymbol{\theta}}$ is used to compute $\hat{\eta}_i \equiv \frac{1}{T}\sum_{t=1}^{T}[y_{it} - \tilde{\tilde{\mathbf{x}}}'_{it}\hat{\boldsymbol{\theta}})]$. The second stage defines $\hat{y}_{it} \equiv y_{it} - \hat{\eta}_i$ and estimates $\hat{\boldsymbol{\theta}}(\tau)$ as:

$$\hat{\boldsymbol{\theta}}(\tau) := \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\rho_\tau(\hat{y}_{it} - \tilde{\tilde{\mathbf{x}}}'_{it}\boldsymbol{\theta})v_{it} \tag{18}$$

In case of endogenous $\mathbf{d}_{it}$ in (6–8), we modify assumption 2 into $u_{it} \perp (\mathbf{x}_{it}, \mathbf{z}_{it}, \eta_i)$. This allows the applicability of Canay's (2011) asymptotic theory and practical two-step procedure. Namely, a $\sqrt{NT}$ consistent estimate of η_i is obtained through a least-squares instrumental variable regression, and then employed for computing $\hat{y}_{it}$. Then, $\hat{y}_{it}$ becomes a dependent variable in (6–8), which is estimated with (Galvao (2011), Chernozhukov and Hansen (2008)) procedure, applied to

$$Q_{NT}(\tau, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) := \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\rho_\tau(\hat{y}_{it} - \mathbf{d}'_{it}\boldsymbol{\alpha} - \mathbf{x}'_{it}\boldsymbol{\beta} - \boldsymbol{\phi}'_{it}\boldsymbol{\gamma})v_{it} \tag{19}$$

As quality-based reimbursement in the U.S. is applied to all Medicare hospitals, endogeneity arises only owing to the lagged dependent variable. So the consistent estimation requires that $u_{it} \perp (\mathbf{d}_{i,t-s}, \eta_i)$, where $\mathbf{d}_{it} = (y_{i,t-1})$, $s = 1, 2, ..., T-1$.

In case of the Japanese length-of-stay reimbursement with voluntary participation in the reform, hospital is assumed to make a decision about introducing PPS, considering the value of its average length of stay in the pre-reform year. So in case of an AR(1) dynamic panel data model with predetermined assignment of the reform r_{it} we assume that $u_{it} \perp (\mathbf{d}_{i,t-s}, \eta_i)$, where $\mathbf{d}_{it} = (y_{i,t-1}, r_{it})$, $s = 1, 2, ..., T-1$. When AR(2) specification becomes necessary for the estimations, we add $y_{i,t-2}$ to the right-hand side of (19) and use third and forth lags as instruments for $y_{i,t-1}$ and $y_{i,t-2}$.

4.3.2 Least squares regression

The analysis is based on autoregressive specification in Hamilton (1994):

$$y_{it} - \mu = \alpha_1(y_{i,t-1} - \mu) + \alpha_2(y_{i,t-1} - \mu)r_{it} + \mathbf{x}'_{it}\boldsymbol{\beta} + \nu_i + \epsilon_{it} \tag{20}$$

The dependent variable, y_{it}, is quality measure or average length of stay. r_{it} is the reform dummy which equals unity if hospital i participates in performance-based payment reform in year t, $\mathbf{x}_{it}$ are exogenous variables (hospital characteristics and annual dummies), ν_i are hospital fixed effects, ϵ_{it} are i.i.d. with zero

mean. The inclusion of the interaction term $(y_{i,t-1}-\mu)r_{it}$ captures the effect of the reform conditional on the pre-reform value of the dependent variable. Note that the conditional quantile regression approach allows direct identification of the differential effect, consequently the interaction term becomes redundant. When μ is significant, there exists an "attraction point": the effect of the reform for hospitals with the pre-reform value of y_{it} greater (smaller) than μ monotonically approaches the effect for hospitals with y_{it} equal to μ "from above" ("from below"). The absence of unit root in the AR(1) process implies $0 < |\alpha_1| < 1$. If an additional condition $0 < |\alpha_1 + \alpha_2| < 1$ holds, then the "attraction point" is the same in the pre-reform and post-reform periods. For convenience we rewrite:

$$y_{it} = \alpha_0 + \alpha_1 y_{i,t-1} + \alpha_2 y_{i,t-1} r_{it} + \alpha_3 r_{it} + \mathbf{x}'_{it}\boldsymbol{\beta} + \nu_i + \epsilon_{it} \tag{21}$$

Equation (21) for the U.S. and for Japan is estimated using Arellano and Bover (1995)/Blundell and Bond (1998) estimator, with robust variance-covariance matrix (Windmeijer (2005)). Since $y_{i,t-1}$ is a factor of $y_{i,t-1}r_{it}$, the interaction term is treated as a predetermined variable. Lagged levels and lagged differences of y_{it} and $y_{i,t-1}r_{it}$ are used as instruments for the differenced equation.

As regards Japanese length-of-stay based reimbursement, hospital is assumed to make a decision about introducing PPS, considering the value of its average length of stay in the pre-reform year. Consequently, r_{it} becomes a predetermined variable, and lagged levels and difference of r_{it} are added to instruments. Arellano and Bond (1991) test does not reject the hypothesis about the absence of order two serial correlation in the first differenced errors in most specifications. In case of the AR(2) specification, we add $y_{i,t-2}$ and $y_{i,t-2}r_{it}$ to the right-hand side of (21), treat $y_{i,t-2}r_{it}$ as a predetermined variable, use third and forth lags as instruments for $y_{i,t-1}$ and $y_{i,t-2}$, and lag of $y_{i,t-2}r_{it}$ as its instrument.

4.3.3 Long-term means

In the least squares approach we set $r = 1$ for the reformed hospitals and adopt the Hamilton (1994) approach to compute the long-term mean μ_r as

$$\mu_r = (\alpha_0 + \alpha_3)/(1 - \alpha_1 - \alpha_2) \tag{22}$$

In case of non-reformed hospitals $r = 0$, so we obtain the conventional form of the long-term mean $\mu_n = \alpha_0/(1 - \alpha_1)$, where subscripts r and n denote reformed and non-reformed hospitals, respectively.

The formulas for the long-term mean for the reformed hospitals in case of AR(2) process modify to $\mu_r = (\alpha_0 + \alpha_3)/(1 - \alpha_1 - \alpha_2 - \kappa_1 - \kappa_2)$ and $\mu_n = \alpha_0/(1 - \alpha_1 - \kappa_1 - \kappa_2)$, where κ_1 and κ_2 are coefficients for $y_{i,t-2}$ and $y_{i,t-2}r_{it}$, respectively.

The quantile regression approach does not employ the interaction terms, so coefficients α_2 and κ_2 are excluded from each corresponding formula.

Both μ_r and μ_n may be contrasted to actual values of thresholds for quality measures in value-based purchasing (as specified in the Federal Register for 2013) or mean length of stay in the Japanese DPC schedule.

4.3.4 Hypotheses

According to our theoretical model, the incentives for changing the performance variables both in the U.S. and Japan depend on the initial values of these variables. In particular, hospitals performing better than the benchmark value are likely not to improve their performance or may even worsen it. Deterioration of the

performance indicator for such hospitals is more plausible for quality measures. Indeed, in case of length-of-stay reimbursement, a hospital's indifference between admitting a new patient or treating a patient who is already in hospital for longer relates not only to per diem pay but also to the capability of sustaining a high rate of bed occupancy (Abe et al. 2005).

Hypothesis I predicts that the U.S. hospitals in the top deciles of quality measures and Japanese hospitals with length of stay less than the cutoff point for period I, respectively, lower quality and increase length of stay. At the same time, hospitals performing worse than the threshold value tend to improve their performance. Accordingly, *Hypothesis II* forecasts that the U.S. hospitals in percentiles 0-50 of quality measures and Japanese hospitals above mean length of stay will, respectively, increase quality and lower length of stay.

Hypothesis III assumes that a move towards "best-practice" rate-setting diminishes the undesired effect of the reform for the best-performing hospitals. A natural experiment with the Japanese pricing schedule allows empirical testing of the hypothesis.

To test *Hypotheses* $I-II$ within the dynamic panel data framework, we examine whether the coefficients $\hat{\alpha}_3$ for the reform dummy r_{it} in quantile regression approach in equation (19)) significantly differs from zero. In case of least squares model we measure $\hat{\delta}_i = \hat{\alpha}_2 \bar{y}_{i,t-1} + \hat{\alpha}_3$ in the AR(1) specification and $\hat{\delta}_k = \hat{\alpha}_2 \bar{y}_{i,t-1} + \hat{\kappa}_2 \bar{y}_{i,t-2} + \hat{\alpha}_3$ in the AR(2) specification, where k indicates a group of hopitals.

In each year we assign each U.S. hospital to a decile group, based on the value of its lagged quality measure. The values of the deciles are used in quantile regressions, and decile groups are taken for the OLS estimations. Regarding the Japanese data, we compute annual cutoff points in the payment schedule – the upper boundaries for period I, II, and III, and take the mean of annual values over the entire period. For each diagnosis-procedure combination and major diagnostic category our computations are based on the empirical distribution of the lagged hospital-level length of stay, weighted by the number of corresponding cases in a hospital. Denote quantiles, corresponding to the mean annual values of the cutoff points for *periods* I, II, and III τ_1, τ_2, and τ_3, respectively. The values of τ_1, τ_2, and τ_3 are used in quantile regressions. In case of OLS models, $k = 1, 2, 3$ indicates percentile groups, corresponding to τ_1, τ_2, and τ_3. Additionally, we look at group 4 of Japanese hospitals, which have lagged length of stay above the length of period III.

A number of issues, relating to the limitations of our approach, become the safeguards in generalising our findings with *Hypotheses* $I - II$. The empirical distribution of performance score is computed at the hospital level in the U.S. and at the patient level in Japan. Since the patient-level data are unavailable, we use the number of patient cases at each Japanese hospital at each diagnoses group as a frequency weight in estimating the empirical percentiles. The approach involves an approximation, which may bias our estimates. This is particularly applicable to the models with diagnosis-procedure combinations, where the number of cases in a hospital may be low and therefore, the variation in the length of stay may be large.

As for contrasting quantile regression and OLS estimates, such comparison can lead only to tentative conclusions. Indeed, $\hat{\alpha}_3$ is the effect of the reform for the whole sample of hospitals, conditional on each corresponding quantile. At the same time $\hat{\delta}_k$ measures the effect of reform for an average hospital in a percentile group, using the coefficients from the mean (OLS) regression. Moreover, the assignment to a percentile group in our OLS-post-estimation is conducted according to a given variable (the lagged outcome). Yet, conditional quantile regression incorporates the influence of all explanatory variables. So $\hat{\alpha}_3$ for each τ and $\hat{\delta}_k$ for comparable groups may be contrasted only in terms of relative values across quantiles/percentiles, but not in term of absolute values.[20]

Finally, the use of dichotomous variables for reform participation allows interpreting only the direction of

[20]The multivariate specification allows only contrasting the value of $\hat{\alpha}_3$ for $\tau = 0.5$ (conditional median regression) with the value $\hat{\delta}_k$ for decile 5 (conditional mean regression).

the effect in each quantile. Indeed, both in the U.S. and Japan benchmarking is established 2-3 years prior to the actual implementation and is based on historic data. So the analysis of the size of the effect would require incorporation of higher-order lags and interaction terms. Yet, the identification of higher-order lags is impossible owing to the short number of post-reform periods in both countries (two time points in the U.S. and at most five time points at the diagnosis-group level in Japan, and additional "loss" of one period due to differencing in dynamic panel estimations).

Concerning *Hypothesis III*, we could not test it by including an interaction term of the reform dummy r_{it} and the dichotomous variable $post_t$ (which would equal unity after the change in benchmarking, i.e. in 2012 and 2013), since it would result in multicollinerarity. Note that although the values may appear insignificant due to large standard errors, the presence of multicollinearity does not prevent consistent estimates of coefficients for the explanatory variables. However, this paper deals with endogenous reforms and we use an instrumental variable approach. Multicollinearity with instrumental variable regression might result in even higher correlation between the fitted values of the endogenous variables at the first stage (Kritzer (1976), Farrar and Glauber (1967)) and, consequently, an upwards bias of the first stage F–statistics, leading to wrong conclusions about the absence of weak instruments (Stock et al. (2002)).

Dividing the data in the pre-change (2007-2011) and post-change periods (2012-2013) would imply using short panels, so the asymptotic results about the consistency of the instrumental variable quantile regression may not hold. Consequently, to assess *Hypothesis III* we exploit the full longitudinal sample but use counterfactual value of τ_1. Instead of setting the upper boundary of *period I* as $\min\{25\text{th percentile}, 0.5\text{mean}\}$ in 2012 and 2013, we compute it as the 25th percentile. *Hypothesis III* predicts smaller extent of the performance deterioration (i.e. smaller/less prevalent increase in the length of stay) at τ_1 (group 1) in these counterfactual estimates if compared to the estimates when the top benchmark group is calculated using the length of *period I* in the changed fee schedule.

5 Data

5.1 Medicare's data on value-based purchasing

The data for clinical process of care and patient experience of care measures come from Hospital Compare (Dec 18, 2014 update). The length of panel covers the period from Jul 2007 to Dec 2013 (releases starting 2009.03.01 onwards).[21] The unit for the time period in our analysis is one fiscal year (2008 to 2013).[22]

Concerning hospital characteristics, we use the data by Hospital Compare on hospital location and ownership (the latter variable as of the 2009 release). The number of hospital beds, share of Medicare's discharges, and the dichotomous variables for urban location (controlling for section 401 hospitals) and teaching status (hospital with non-zero residents) are taken from Medicare's Impact Files. The data on casemix index come from the Final Rules for each fiscal year.

Combining Hospital Compare and Medicare's data files, we select Medicare's acute-care hospitals, as value-based purchasing applies exclusively to this subgroup.

In estimating monetary gain of increasing quality under the reform we use the Final Rules and Medicare provider utilization and payment data (inpatient care) on the number of Medicare's discharges and hospital-level average reimbursement for each discharge for fiscal years 2011–2013.

[21] We begin with the first period that has available coded data for the quality measures.

[22] The data for most measures in Hospital Compare are based on the results of the 12-month surveys. The values of HCAHPS measures are updated quarterly, adding the data for the new quarter available and excluding the data for the corresponding first quarter. However, clinical process of care measures are updated annually. Similarly, hospital control variables are reported on fiscal year basis.

Although the financial redistribution of funds within value-based purchasing came in force in the fiscal year 2013, the rewards schedule was established in May 2011 based on historic data. So each hospital realized its position relative to the empirical distribution of performance measures upon the announcement of the benchmark, threshold and floor figures, as well as the slope for the linear exchange function. Moreover, hospitals were aware that the results of the fiscal year 2012 survey (which was launched in October 2011 and included explicit calculations of both achievement and improvement points) would be used in price-setting for 2014 onwards.[23] Consequently, it is plausible to assume that incentives within value-based purchasing apply to survey-participant hospitals starting from fiscal year 2012.

Accordingly, our dichotomous variable for reform participation equals unity in fiscal years 2012 or 2013 if a hospital is listed as a value-based purchasing hospital in the survey database for the corresponding year (Table 6).

5.2 Japanese data on length-of-stay performance

The analysis employs an administrative database from Japan's Ministry of Health, Labor, and Welfare (September 5, 2014) on annual hospital-level aggregated information for patients, discharged in July-October 2005, July-December 2006-2010, July 2011-March 2012, April 2012-March 2013, and April 2013-March 2014. The data are voluntarily sent to MHLW by hospitals, which plan to join the PPS reform. Hospitals may join the PPS reform after the trial period (commonly after two years), may postpone the decision and keep submitting the data to the MHLW, or may choose to never join the reform and discontinue sending their data. Merging the MHLW's annual files by hospital name (checking for any change of name due to restructuring, mergers, and closures), we construct an unbalanced panel of 1849 hospitals, which have submitted data to MHLW since 2005.

Given data availability, we conduct the analysis at the level of 10-digit code diagnosis-procedure combinations (fiscal years 2007-2013). Overall, there are 743 such DPC groups, existing within the analyzed period. Yet, only 175 of them have enough cases for longitudinal estimates at the hospital level (we use the criterion for the DPC group present in more than 100 hospitals in the unbalanced panel). Finally, we exploit the MHLW's data for the DPC price schedules (revisions of 2006, 2008, 2010 and 2012), selecting DPCs with flat rates at the 10-digit level. This enables merging the length of stay and price data for 32 ten-digit DPC groups.

Additionally, we do estimations at the aggregated level of MDCs in 2007-2013. The panels are unbalanced, but 84%-94% of hospitals have observations for corresponding MDC all the years. Adding earlier years (2005 and 2006) would decrease the percent of such hospitals to 20%, since 76% of hospitals joined the MHLW data base only in 2007. Table 2 gives the summary statistics at the MDC and hospital level, and the list of 175 DPCs used in this paper is presented in the Appendix C (Table D.7).

It should be noted that 16 MDCs existed in Japan in the pre-2008 period. In 2008 the 16th MDC, which encompassed four clinical entities,[24] was subdivided into three categories: "Trauma, burns, poison" (new MDC 16); "Mental diseases and disorders" (new MDC 17), and "Miscellaneous" (new MDC 18). So our econometric analysis with MDC16 uses the data only for 2008-2013. (Aggregating/disaggregating of certain diagnoses in Japanese MDCs relative to ICD-10 is explained in Appendix B.)

Using the nationwide data on teaching hospitals from the Japan Residency Matching Program (2003–2013) we discover that having non-zero residents in a given year is related to hospital's productive efficiency, which is strongly correlated with length of stay (Besstremyannaya (2015), Besstremyannaya (2011)). Con-

[23] More precisely, Final Rule for FY 2014 is based on the May 1, 2011 – Dec 31, 2011 survey.

[24] trauma, burns, poison and the toxic effect of drugs; mental diseases and disorders; diseases and disorders of systemic infection; and miscellaneous (Kuwabara et al. (2008))

sequently, using a control variable of a hospital with non-zero residents would lead to endogeneity. Instead, we exploit a dichotomous variable for university hospitals, and in this way we control for a combination of teaching and research activities.

The data from the Japan Council for Quality Health Care (2014) enables constructing a time-varying dichotomous variable as a proxy for hospital quality, which equals unity if the hospital is given accreditation by the beginning of the corresponding financial year.

The dichotomous variables for university and emergency hospitals come from the 2014 online version of the Handbook of Hospitals (Byouin yoran). The MHLW (2012d) data are employed to create a time-varying dichotomous variable with unity value for hospitals, which received the status of designated local hospital (and hence, subsidy per each admission) by the beginning of the financial year. Since ownership is shown to be a significant determinant of length of stay in Japan (Kuwabara et al. (2011), Kuwabara et al. (2006)), we construct dichotomous variable for public hospitals.

Concerning hospital size, the MHLW database reports the number of DPC beds only in 2010-2013. Therefore, we use the hospital-level share of DPCs in the national list (available for each year) as a proxy for the share of DPC beds in the total number of acute-care beds: pairwise correlations between the two variables in 2010-2013 range from 0.9439 to 0.9509.

Table 1: **Descriptive statistics for the unbalanced panel of Medicare hospitals in 2008-2013**

Variable	Definition	Obs	Mean	St.Dev	Min	Max
Patient experience of care measures						
Comp-1-ap	Nurses always communicated well	19184	69.99	7.40	7	100
Comp-2-ap	Doctors always communicated well	19184	75.6	6.25	32	100
Comp-3-ap	Patients always received help as soon as they wanted	19184	79.87	5.39	30	100
Comp-4-ap	Pain was always well controlled	19182	63.06	8.83	18	100
Comp-5-ap	Staff always gave explanation about medicines	19175	68.99	5.64	7	100
Comp-6-yp	Yes, staff did give patients discharge information	19174	60.36	6.67	12	100
Clean-hsp-ap	Room was always clean	19175	82.32	5.10	35	100
Quiet-hsp-ap	Hospital always quiet at night	19183	57.45	10.53	19	100
Hrecomddy	Patients who would definitely recommend the hospital	19183	68.91	10.08	2	100
Hsp-rating-910	Patients who gave hospital a rating of 9 or 10 (high)	19181	66.64	9.27	24	100
Clinical process of care measures						
AMI-8a	Primary PCI received within 90 minutes of hospital arrival (AMI)	8342	70.44	32.82	0	99
HF-1	Discharge instructions (heart failure)	18275	75	29.59	0	99
PN-3b	Blood cultures performed in the emergency department prior to initial antibiotic received in hospital (pneumonia)	18111	82.49	29.87	10	99
PN-6	Initial antibiotic selection for CAP in immunocompetent patient (pneumonia)	18348	84.46	23.86	8	99
SCIP-Card2	Surgery patients on beta-blocker therapy prior to arrival who received a beta-blocker during the perioperative period	13893	75.6	34.04	0	99
SCIP-Inf1	Prophylactic antibiotic received within 1 hour prior to surgical incision	17909	78.52	33.73	0	99
SCIP-Inf2	Prophylactic antibiotic selection for surgical patients	17916	79.09	34.66	0	99
SCIP-Inf3	Prophylactic antibiotics discontinued within 24 hours after surgery end time	17862	83.99	26.7	0	99
SCIP-Inf4	Cardiac surgery patients with controlled 6 A.M. postoperative blood glucose	6792	86.85	22.15	10	99
SCIP-VTE2	Surgery patients who received appropriate venous thromboembolism prophylaxis within 24 hours prior to surgery to 24 hours after surgery	18033	81.69	27.98	0	99
Reform dummy						
VBP	=1 in 2012 and/or 2013 if value-based purchasing hospital in the corresponding fiscal year	19184	.14	.347	0	1
Hospital characteristics						
public	=1 if government's hospital	19184	.18	.38	0	1
emergency	=1 if emergency hospital	19184	.95	.22	0	1
urban	=1 if urban hospital	19184	.725	.45	0	1
teaching	=1 if teaching hospital (i.e. has residents in a given year)	19184	.315	.46	0	1
casemix	transfer-adjusted casemix index	19184	1.43	.30	.47	4.81
beds	number of beds	19184	192.61	177.27	2	1928
medicare_share	share of Medicare cases	19184	.47	.15	.001	1

Notes: Clean-hsp-ap and Quiet-hsp-ap albeit measured separately, are regarded as one measure "Cleanliness and quietness of hospital environment" in the Final Rule. Hrecomddy is not listed in the Final Rule, yet, we analyze it since it relates to overall rating of hospital. We do not analyze the dynamics of two clinical process of care measures from the Final Rule: AMI-7a (Fibrinolytic therapy received within 30 Minutes of hospital arrival), owing to non-availability of its 2013 data, and SCIP-VTE1 (Surgery patients with recommended venous thromboembolism prophylaxis ordered), which was discontinued in 2013. Hospitals with fewer than 10 patients in the survey for the corresponding clinical process of care measure are not included, since they are not scored according to FY 2013 Final Rule. N.a. = non-applicable, since floor is employed in estimating the scores only for HCAHPS measures. *Government* includes federal, state or local government and hospital district or authority. Section 401 hospitals are treated as rural hospitals.

Source: CMS FY 2013 final rule. Federal Register, Vol.76, No.88, May 6, 2011, Tables 4 and 9.

Table 2: **Descriptive statistics for the unbalanced panel of Japanese hospitals in 2007-2013**

Variable	Definition	Obs	Mean	St.Dev	Min	Max
Length of stay (los) in days for major diagnostic category (MDC)						
los_MDC1	Nervous system diseases	10845	20.41	5.62	3.39	61.06
los_MDC2	Eye system diseases	7270	5.76	2.54	2	18.89
los_MDC3	Ear, nose, mouth, and throat system diseases	10269	8	3.95	2	41.2
los_MDC4	Respiratory system diseases	11010	17.74	4.77	2.7	56.85
los_MDC5	Circulatory system diseases	10672	14.82	4.91	2.28	47.82
los_MDC6	Alimentary, liver, biliary-tree, and pancreas diseases	10903	13.74	2.99	2	38.76
los_MDC7	Musculoskeletal and connective tissue system diseases	10717	19.8	5.3	2.82	74.65
los_MDC8	Skin and subcutaneous tissue diseases	9056	13.19	4.62	3.03	58.91
los_MDC9	Breast system diseases	7003	11.56	4.68	2	52.62
los_MDC10	Endocrine, nutritional, and metabolic system diseases	10842	16.06	4.29	3.36	51.29
los_MDC11	Kidney, urinary tract, and male reproductive system diseases	10724	14.42	4.48	3.78	58.6
los_MDC12	Female reproductive system and puerperal diseases, abnormal pregnancy, and abnormal labor diseases	6532	10.91	3.21	2.09	74.67
los_MDC13	Blood and blood forming organs an	9358	23.38	8.47	3.38	66.03
los_MDC14	Newborn and other neonates, congenital anomalies diseases	5749	11.17	5.19	2.14	36.73
los_MDC15	Pediatric diseases	9892	8.05	2.6	2.32	31.32
los_MDC16	Trauma, burns, poison	9557	18.02	5.56	2.8	55.58
Reform variables						
PPS	=1 if introduced PPS by corresponding fiscal year	11422	.72	.45	0	1
Hospital characteristics						
public	=1 if public hospital	11422	.24	.43	0	1
emergency	=1 if emergency hospital	11422	.86	.35	0	1
urban	=1 if urban hospital	11422	.98	.14	0	1
university	=1 if university hospital	11422	.07	.26	0	1
DPC beds	number of hospital beds, for which diagnoses are coded (and reimbursed according to PPS under PPS)	6811	292.82	209.22	0	1445
share_DPClist	share of diagnosis procedure combinations, treated by hospital in a given year, in the national list of diagnosis-procedure combinations	11422	.2	.12	.01	.58
designated	=1 if given the status of designated hospital by corresponding fiscal year	11422	.22	.41	0	1
quality	=1 if given independent third-party accreditation by the Japan Council for Quality Health Care by the corresponding fiscal year	11422	.59	.49	0	1

Notes: 1) The Japanese MDC6 encompasses MDC6 and MDC7 in ICD-10, MDC11 incorporates MDC11 and MDC12 in ICD-10, MDC12 combines MDC13 and MDC14 in ICD-10, MDC13 includes MDC16 and MDC17 in ICD-10. At the same time, MDC9 in ICD-10 is disaggregated into the Japanese MDC8 and MDC9. MDC16 is distinguished as a group only since 2008. 2) Number of DPC beds is available only for 2010-2013. 3)Prefecture grants the status of designated hospital and financial support of 10,000 yen per each admission to municipal or regional hospital which satisfies the following requirements: has over 200 beds; the share of patients referred from other facilities is over 60%; shares its beds and expensive equipment (such as MRI, CT scanner) with other hospitals; educates health care officials; has emergency status. 4) The third-party accreditation is started in Japan in 1997, and is granted to hospitals that fulfill seven standards: mission, policy, organization and planning; community needs; medical care and medical care support systems; nursing care; patient satisfaction and safety; administration; specific standard for rehabilitation and psychiatric hospitals (Hirose et al. (2003)). 5) English names of MDCs in Ministry of Health, Labor and Welfare (2014a) are adopted from Hayashida et al. (2009), Kuwabara et al. (2008) and Ishikawa et al. (2005). 6)Public hospitals are national (*kokuritsu*), prefectural (*kenritsu*, *douritsu*, *furitsu*), city (*shimin*, *shiritsu*), town (*chouritsu*), village (*sonritsu*), municipal (*kouritsu*) hospitals, and hospitals in National Health Insurance system (*kokuho*) and the system for health care of workers (*roudousha kenkou fukushi kikou*).

6 Results

6.1 Quality

Since the pricing schedule relates to all hospitals in the annual empirical samples, this section presents the results with longitudinal data, which may lack observations in certain years. Overall, the panels are unbalanced but 85–93% of hospitals would have observations in each year. As robustness check, we conducted analysis with balanced panels and discovered similar distribution of the dependent variables and negligible difference in the values for the coefficients for the explanatory variables.

Identification condition for AR(1) process and Arellano and Bond (1991) test not rejecting the hypothesis about the absence of order two serial correlation in the first differenced errors hold for nine HCAHPS measures (exception is Comp-6yp) and for six clinical process of care measures: AMI-8a, HF-1, SCIP-INf1, SCIP-Inf3, SCIP-Inf4, SCIP-VTE2. Owing to unavailability of longer time-series for post-reform data, we cannot estimate higher order lags and limit our analysis to the above 15 measures.

The results of quantile regression estimates demonstrate that the fixed effect model is preferred to the random effects model for all analyzed HCAHPS measures (with the exception of Comp-5-AP) and the lagged dependent variable is significant. Concerning clinical process of care measures, in 83% of our models the preferred model is with fixed effects and the lagged dependent variable is significant. Similarly, analysis with the OLS dynamic panel data model reveals that the coefficient for the lagged dependent variable is significant for all but one analyzed HCAHPS measures and clinical process of care measures, indicating the presence of "habit-formation".[25]

The effect of VBP for HCAHPS measures is heterogeneous across quantiles and is inversely related to the value of quantile, as predicted by our $Hypotheses\ I - II$. In other words, improvement of quality measures owing to the reform is observed in quantiles 0.1-0.4, while quantiles 0.6-0.95 demonstrate a negative effect of the reform. Similarly, the reform has a positive effect for percentile groups below the mean and a negative effect for percentiles above the mean, according to OLS models (Table 3, Figure 2 and Table D.3).

[25] The interaction term is negatively significant and the reform dummy has a positive estimated coefficient in the OLS models. The sum of coefficients for the lagged dependent variable and the interaction term has an absolute value less than unity and is significant for all measures, with the exception of $Hrecomddy$.

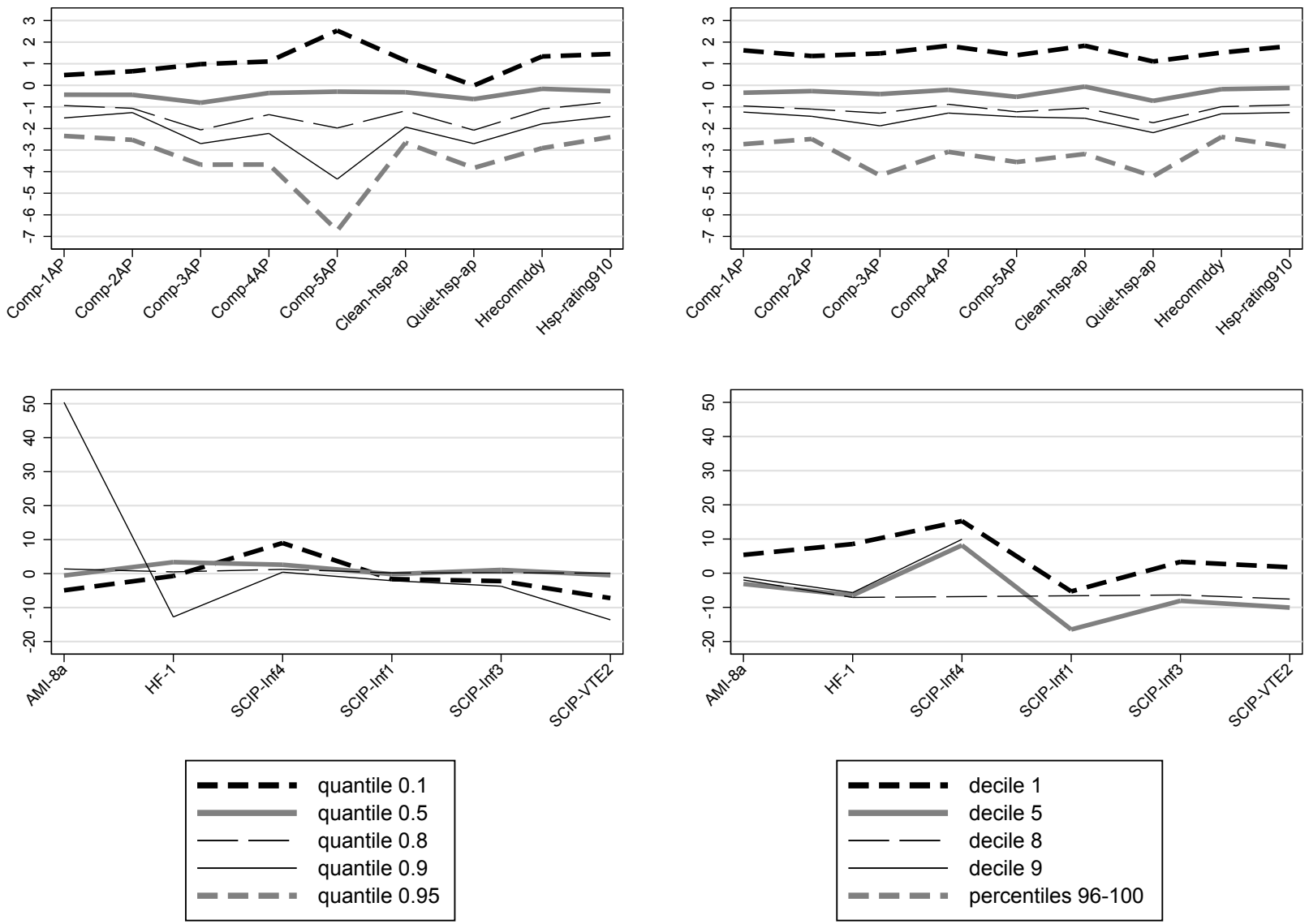

Figure 2: The effect of value-based purchasing for quality measures (left: quantile regression, right: OLS models)

Notes: Models for clinical process of care measures generally do not converge for $\tau = 0.95$, and estimates for deciles9 and percentiles 96-100 for most measures are not available, so the results for top quantiles may be contrasted at $\tau = 0.8$ and decile 0.8.

Table 3: **Coefficient for the value-based purchasing dummy for quality measures in dynamic panel quantile regression**

	HCAHPS									**Clinical process of care**					
	Comp-1ap	Comp-2ap	Comp-3ap	Comp-4ap	Comp-5ap	Clean-hsp-ap	Quiet-hsp-ap	Hrecomddy	Hsp-rating910	AMI-8a	HF-1	SCIP-Inf1	SCIP-Inf3	SCIP-Inf4	SCIP-VTE2
$\tau = 0.1$	0.477***	0.65***	0.983***	1.106***	2.537***	1.144***	-0.011	1.338***	1.45***	-4.933	-0.704	-1.62**	-2.201***	9.002*	-7.21
	(0.186)	(0.251)	(0.387)	(0.277)	(0.143)	(0.313)	(0.3)	(0.358)	(0.298)	(164.49)	(5.918)	(0.803)	(0.836)	(5.414)	(151630)
$\tau = 0.2$	0.144	0.331***	0.258	0.417*	0.911***	0.464*	0.101	0.841***	0.603***	2.998*	4.638	-0.704	4.376	0.462	0.748
	(0.157)	(0.132)	(0.277)	(0.224)	(0.098)	(0.267)	(0.238)	(0.243)	(0.232)	(1.604)	(7.181)	(0.46)	(269.61)	(0.535)	(97.087)
$\tau = 0.3$	-0.148	-0.062	-0.275	0.109	0.581***	0.162	-0.236	0.359**	0.306	-0.653	1.353	-0.557	0.601	0.373	-0.216
	(0.149)	(0.114)	(0.213)	(0.183)	(0.11)	(0.189)	(0.179)	(0.203)	(0.211)	(31.404)	(1.437)	(1.788)	(1.999)	(8.398)	(1.81)
$\tau = 0.4$	-0.337***	-0.245**	-0.498***	-0.136	0.085	-0.204	-0.31	0.118	-0.09	-0.391	0.412	-0.142	0.588	0.55	-0.092
	(0.124)	(0.114)	(0.206)	(0.145)	(0.11)	(0.182)	(0.193)	(0.182)	(0.186)	(2.843)	(0.948)	(0.839)	(1.026)	(2.431)	(1.385)
$\tau = 0.5$	-0.433***	-0.437***	-0.806***	-0.355***	-0.292***	-0.32**	-0.636***	-0.161	-0.266	-0.569	3.374***	-0.17	1.088*	2.604**	-0.469
	(0.12)	(0.116)	(0.192)	(0.152)	(0.105)	(0.173)	(0.194)	(0.175)	(0.175)	(3.151)	(1.012)	(0.948)	(0.588)	(1.26)	(0.632)
$\tau = 0.6$	-0.565***	-0.687***	-1.08***	-0.661***	-0.558***	-0.444**	-0.995***	-0.369**	-0.451***	0.34	0.771	-1.012	0.4	1.167	-0.659
	(0.14)	(0.116)	(0.191)	(0.148)	(0.107)	(0.193)	(0.233)	(0.177)	(0.171)	(1.709)	(0.834)	(0.718)	(0.309)	(1.303)	(0.617)
$\tau = 0.7$	-0.829***	-0.787***	-1.318***	-0.916***	-1.085***	-0.811***	-1.617***	-0.645***	-0.658***	-0.326	0.396	-0.888	-0.064	0.773	0.033
	(0.131)	(0.126)	(0.263)	(0.154)	(0.117)	(0.224)	(0.251)	(0.215)	(0.19)	(1.903)	(0.326)	(0.667)	(0.533)	(0.644)	(0.248)
$\tau = 0.8$	-0.937***	-1.063***	-2.063***	-1.356***	-1.979***	-1.178***	-2.077***	-1.094***	-0.758***	1.365**	0.507*	0.2	0.25	1.249***	0.132
	(0.127)	(0.123)	(0.238)	(0.187)	(0.155)	(0.253)	(0.224)	(0.203)	(0.173)	(0.609)	(0.277)	(0.288)	(0.213)	(0.503)	(0.192)
$\tau = 0.9$	-1.51***	-1.263***	-2.698***	-2.227***	-4.344***	-1.934***	-2.704***	-1.782***	-1.44***	50.402***	-12.755***	-2.093	-3.713***	0.418	-13.595*
	(0.212)	(0.128)	(0.411)	(0.266)	(0.159)	(0.254)	(0.292)	(0.334)	(0.315)	(4.454)	(4.858)	(1.606)	(1.131)	(1.079)	(7.743)
$\tau = 0.95$	-2.346***	-2.525***	-3.673***	-3.665***	-6.728***	-2.658***	-3.826***	-2.91***	-2.394***	3.821	3.934	2.12	-1.164	1.362	5.339
	(0.377)	(0.342)	(0.377)	(0.553)	(0.253)	(0.459)	(0.542)	(0.473)	(0.47)	(20.456)	(2.897)	(5.01)	(3.902)	(5.566)	(5.061)

Note: Robust standard errors in parentheses. The table reports the results of the FE model, with exception of $Comp-5ap$. Similarly, RE models are used for three clinical process of care measures for $\tau = 0.1$ and $\tau = 0.2$, one measure for $\tau = 0.6$ and 3 measures for $\tau = 0.7$, according to the results of the Hausman test and/or to secure convergence. Models for clinical process of care measures generally do not converge for $\tau = 0.95$. Owing to extremely skewed data, bound values for l.score coefficient (set at 0.01 and 0.99) remain for 13% of models.

The estimates of the long-term means $\hat{\mu}_r^{OLS}$ and $\hat{\mu}_n^{OLS}$ for the reformed and non-reformed hospitals are significant for all HCAHPS measures. According to the results of the Wald test, $\hat{\mu}_r^{OLS} = \hat{\mu}_n^{OLS}$. Similarly, in case of quantile regression for the 50th percentile, $\hat{\mu}_r^{QR} = \hat{\mu}_n^{QR}$. Although $\hat{\mu}_r$ and $\hat{\mu}_n$ are slightly higher than threshold values in the FY 2013 Medicare Final rule,[26] the statistical difference between $\hat{\mu}_r$ (or $\hat{\mu}_n$) and the threshold is insignificant.

The similarity of the results for quantile regressions and OLS models in terms of the reform effect and the values of the long-term means may be explained by the fact that the distribution of the dependent variable is not skewed and the errors in the OLS model are close to normal (Table D.1).

At the same time the distribution of each clinical process of care measure is extremely skewed: a large share of hospitals belongs to the bottom decile and very few are in the top deciles.[27] The distribution of the error term in the OLS regressions is bimodal, indicating decrease of scores in the bottom decile and increase in the top deciles. Indeed, the quantile regressions show that the reform dummy generally has a negative coefficient for $\tau = 0.1$ and positive coefficients for $\tau > 0.8$. The absence of convergence can be seen from the insignificant values of the long-term values of $\hat{\mu}_r$ (or $\hat{\mu}_n$) for a number of clinical process of care measures.

However, the negative coefficient in the bottom decile does not necessarily relate to quality deterioration. Indeed, participation in the VBP resulted in better reporting and expanded patient samples. These larger samples, which particularly apply to the bottom decile, make it possible to differentiate between extremely low scores and to reveal quality more accurately. The positive coefficient of the reform dummy for some measures in top deciles is due to negative interdependence between the scores for the measures.[28] Consequently, decrease in efforts for sustaining quality for a certain measure, where a hospital already shows good performance, may result in releasing extra labor or capital resources, which may be spent for increasing the quality of other measures. Moreover, the best hospitals often do not participate in the VBP, and therefore, the coefficients for the reform dummies in the quantile regressions have large standard errors and the effect in the OLS regressions could not be identified. Overall, owing to the skewness of data and short data series for clinical process of care measures, we can only tentatively assess our hypothesis about the link between quality and performance.

Regarding HCAHPS measures, the results of the estimates with quantile regression and linear dynamic panel data model offer persuasive evidence for non-rejection of *Hypotheses* $I - II$. Indeed, according to the VBP payment schedule, the adjustment coefficient is expected to equal unity for hospitals above the 50th percentile of total performance score. Under an assumption that scores do not differ appreciably across measures within each hospital, hospitals above the 50th percentile of each measure may not have incentives to improve their performance if they only want to remain budget neutral. Moreover, the closer the hospital is to the benchmark value (or if the hospital is above the benchmark), the harder it is to improve the quality measures. So the quality in top deciles does not increase or may deteriorate, and the scope of quality decrease is positively related to the baseline achievement of a hospital.

[26] This fact may indicate the potential skewness of data in the baseline period, which were used for calculations of VBP target values.

[27] The skewness does not disappear if we take the fully balanced panel.

[28] While with HSAHPS measures there is a significant positive correlation between the scores for the overall sample and in each decile, the scores for clinical process of care measures have a weak negative correlation at the top deciles.

6.2 Length of stay and change in benchmarking

The results of our estimates reveal strong evidence in support of *Hypotheses* $I - II$. Regarding quantile regressions, the coefficient for the prospective payment reform at τ_1 is positive and significant for five out of fourteen estimated models with major diagnostic categories, and insignificant in the remaining models. In OLS estimations, $\hat{\delta}_1$ is positive and significant for eleven major diagnostic categories and insignificant for the remaining five. Concerning *Hypothesis II*, the reform coefficient at τ_3 in quantile regressions is negatively significant for seven major diagnostic categories, and insignificant for others. $\hat{\delta}_3$ and $\hat{\delta}_4$ are negatively significant in, respectively, 81.2% and 93.8% percent of OLS models for major diagnostic categories, and insignificant in the remaining models (Figure 3, Table 4).

At the level of diagnosis-procedure combinations we rely mainly on the OLS estimates. Indeed, the small values of τ_1 (often close or even below 0.1) and high values of τ_3 (often above 0.9) would require very large data samples for quantile regression estimates (e.g. at least 1000 observations in the unbalanced panel, so that the number of observations in the quantile group were at least 100 hospitals).[29] The OLS estimates reveal that $\hat{\delta}_1$ is either positive or insignificant for all analyzed diagnoses and positively significant for 60.6% of diagnoses, so the *Hypothesis I* is never rejected. $\hat{\delta}_3$ and $\hat{\delta}_4$ are negatively significant in 61.7% and 77.7% of models for diagnosis-procedure combinations. *Hypothesis II* may be rejected in only 1 out of 175 models (Table 5).

Regarding *Hypothesis III* the positive effect on the length of stay at τ_1 (and group 1) in the counterfactual benchmark estimations is generally smaller than the actual effect of the reform in all the models, where the effect is significant. The number of models with positively significant effect remains the same. However, for two MDCs in quantile regression the effect becomes negatively significant in counterfactual estimates. Similarly, the analysis at the level of diagnosis-procedure combinations shows that the coefficient in counterfactual estimates is smaller in 97.2% of models, equal to the coefficient in the actual estimates in 1.9% of models and larger in only 1 model out of 106 (where it was positively significant). The significance with counterfactual estimates is lost in five models out of 106.

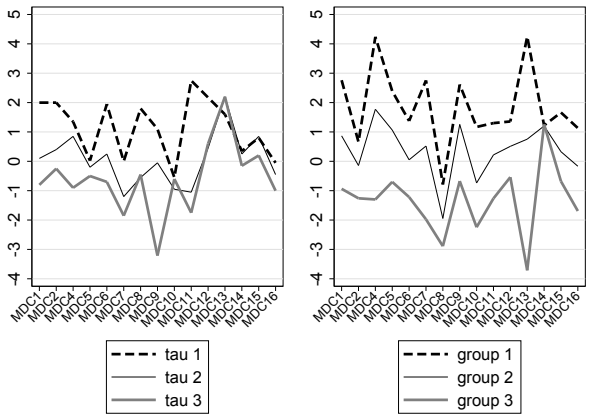

Figure 3: The effect of inpatient prospective payment for length of stay (left: quantile regression, right: OLS models)

Notes: τ_1, τ_2, and τ_3 are percentiles for, respectively, lengths of *period* I, II and III for lagged $\mathbf{y}$ in the corresponding years (average of annual point estimates over the time period); groups $k = 1, ..., 3$ include hospitals, who have lagged length of stay not exceeding the length of period I, II and III, respectively. MDC3 is omitted, since it would require AR(3) estimations, which could not be done using data available.

[29] This criterion is satisfied for only 17 diagnosis-procedure combinations.

Table 4: **Coefficients for the prospective payment system dummy for length of stay and values of percentile points in dynamic panel quantile regression**

	MDC1	MDC2	MDC4	MDC5	MDC6	MDC7	MDC8	MDC9	MDC10	MDC11	MDC12	MDC13	MDC14	MDC15	MDC16
Coefficients at percentiles															
τ_1(new benchmark)	1.999***	1.999	1.349	0.024	1.949	0.001	1.799***	1.099***	-0.551	2.749***	n.a.	1.599*	0.349	0.799	-0.051
	(0.43)	(5.837)	(0.936)	(0.553)	(3.342)	(0.367)	(0.382)	(0.351)	(0.356)	(0.578)	n.a.	(0.872)	(0.515)	(0.529)	(0.406)
τ_1(counterfactual old benchmark)	1.051	1.851	0.701***	-0.201	1.001	-1.001***	0.899	1.751	-0.851***	3.001***	1.601	1.901***	0.501	0.999***	-0.351
	(0.264)	(2.126)	(0.342)	(0.205)	(5.526)	(0.366)	(1.181)	(1.599)	(0.387)	(0.851)	(1.451)	(0.795)	(0.369)	(0.249)	(0.235)
τ_2	0.099	0.399	0.849	-0.201	0.249	-1.201	-0.551	-0.051	-0.951	-1.051	0.449	2.199	0.249	0.849	-0.451
	(0.382)	(0.186)	(0.259)	(5.502)	(0.176)	(0.286)	(1.223)	(0.642)	(0.245)	(0.39)	(0.312)	(2.600)	(0.45)	(0.144)	(0.23)
τ_3	-0.801***	-0.251	-0.901**	-0.501	-0.701***	-1.851***	-0.451	-3.211***	-0.601	-1.751***	0.599	2.199	-0.151	0.199	-1.001*
	(0.204)	(0.431)	(0.399)	(0.662)	(0.294)	(0.73)	(0.305)	(0.475)	(0.654)	(0.467)	(0.429)	(1.595)	(0.53)	(0.214)	(0.512)
Values of percentile points															
τ_1(new benchmark)	0.151	0.323	0.086	0.075	0.110	0.123	0.09	0.118	0.102	0.090	0.125	0.286	0.403	0.089	0.114
τ_1(counterfactual old benchmark)	0.282	0.381	0.198	0.161	0.277	0.248	0.188	0.213	0.171	0.207	0.284	0.414	0.437	0.195	0.247
τ_2	0.499	0.652	0.426	0.399	0.517	0.474	0.487	0.486	0.478	0.457	0.555	0.657	0.714	0.516	0.492
τ_3	0.940	0.971	0.912	0.869	0.942	0.935	0.919	0.895	0.931	0.896	0.942	0.964	0.977	0.921	0.923

Note: Robust standard errors in parentheses. MDC3 is omitted, since it would require AR(3) estimations, which could not be done using data available. For each major diagnostic category hospital-level length of stay is weighted by the number of patient cases in computing the empirical nationwide distribution. "New benchmark" estimates the upper boundary of period I as $\min\{25\text{th percentile}, 0.5\text{mean}\}$ in 2012 and 2013. "Counterfactual old benchmark" computes τ_1 as the 25th percentile of length of stay in all years. The values of τ_4, which is a percentile point corresponding to hospitals with length of stay above period III, is unity, therefore, computations with quantile regressions could not be identified. The table reports the results of the FE model, with exception of MDC1 at τ_1, MDC8, MDC15 at τ_2, and MDC1, MDC8, MDC9, MDC14, MDC15 at τ_3. Still, bound value for lagged length of stay coefficient (set at 0.01 and 0.99) remains for 1 model (out of 45 estimated) and no convergence was achieved at MDC12 at τ_2. No convergence was achieved at τ_1 (new benchmark) with MDC12.

Table 5: **The effects of PPS on average length of stay at the diagnosis-procedure combination level**

	All MDCs	MDC1	MDC2	MDC3	MDC4	MDC5	MDC6	MDC7	MDC8	MDC9	MDC10	MDC11	MDC12	MDC13	MDC14	MDC16
$\delta_1 > 0$ (new benchmark)	106	10	4	7	8	8	28	3	1	2	3	8	14	4	1	5
$\delta_1 > 0$ (counterfactual old benchmark)	101	9	3	7	8	8	26	3	1	2	2	8	14	4	1	5
$\delta_3 < 0$	108	9	8	5	6	7	32	2	1	1	2	13	9	3	2	8
$\delta_4 < 0$	136	12	8	9	9	10	38	3	1	2	3	15	13	3	2	8
Total analyzed DPCs	175	19	9	10	10	11	44	3	1	2	3	17	19	5	2	20

Note: For each MDC the table lists the number of analyzed 10-digit DPCs; and the number of DPCs with corresponding statistically significant effects. See lists with the codes of 158 DPCs in the Appendix.

With exception of one DPC group, the mean long-term value of length of stay, to which the reformed hospitals converge, is larger than the value of period *II* in the MHLW's schedule. Similarly to the findings with the U.S. data, the differences between $\hat{\mu}_r$ and the threshold (here defined as the final day for period *II*, i.e. the empirical mean) are statistically insignificant for most of analyzed DPCs. Yet, in case of non-reformed hospitals $\hat{\mu}_n$ is different from the threshold in almost 50% of models.

6.2.1 Internal validity

To assess internal validity of the estimates with Hospital Compare data we use the subsample of hospitals with a share of Medicare revenues above 50% and observe higher values for the adverse effects from HSAHPS measures.

Concerning the analysis with length-of-stay performance in Japan, submission of data before reform participation was voluntary. Nonetheless, the Japanese hospitals joining the reform are subject to the rates, related to the percentiles of the length of stay in the sample of data-submitting hospitals. So the effect for hospitals in percentiles 0-25 (if they join the reform) would be opposite to the effect for hospitals above the mean, regardless of how the particular group of data-submitting hospitals compares to all national hospitals.

However, the scale of the effect may depend on the annual subsamples of all data-submitting hospitals. This is why we use the full sample of hospitals in 2005-2013 and compare the effects in the two largest hospital cohorts: submitting data since 2007 (699 hospitals) and since 2006 (371 hospitals). The hospitals from the first cohort are distributed in the highest percentiles of the first quartile of the length of stay in the first year of submitting the data. On the contrary, hospitals in the second cohort are distributed more uniformly within the first quartile. This results in the higher prevalence of the effects, forecast by *Hypothesis I* in the second cohort.

The major limitation of our analyses is the lack of patient data. Consequently, we cannot control for individual socio-demographic characteristics, which may influence treatment patterns and bias our estimates. Regarding the Japanese data, the lack of patient data does not let us use the actual empirical distribution of length of stay. Finally, our attempt to reconstruct the distribution using hospital-level averages, weighted by the number of patient cases at the MDC or DPC level, may be biased for hospitals with a small number of patients.

7 Discussion

The empirical analysis in this paper confirms the presence of adverse effects of performance-based reimbursement with linear or stepwise exchange function, linking performance to payment. It should be noted that the hypotheses concerning differential effects for subgroups of hospitals are essentially built on the fact that there is a certain population of hospitals to which the rates apply (Monrad Aas 1995).[30] So the threshold and/or benchmark value in the national schedule may be worse than the value in a given hospital. Therefore, quality or length-of-stay performance based reimbursement with benchmarking becomes a cause of undesired effects.

In this regard, we propose linking reimbursement of hospitals above a decile-related benchmark to their improvement or establishing benchmark at the value of the best performing hospital. A gradual increase of the values of benchmark and threshold for each quality measure [31] may be viewed as a step towards "best-practice" benchmarking in the U.S. Similarly, Japanese MHLW's (2012a) decrease of *period I* to one

[30] These are at least half of the top decile of quality measure in the U.S. and the bottom quartile of length of stay in Japan.
[31] See final rules for Medicare's value-based purchasing in FY 2014 and 2015.

day for 22 DPCs with high medical costs may be regarded as a move towards "best practice" rate-setting.[32] "Episode-based" payment, rewarding a hospital for treating each patient case when the corresponding criteria were satisfied, may offer an alternative solution to the unwanted effects of performance-based reimbursement (Werner and Dudley (2012), Rosenthal (2008)).

We also suggest that along with examining total performance score, more attention be paid to the dynamics of each quality measure. Our findings indicate that a "habit-formation" model and quality-based reimbursement may not be applicable to certain clinical process of care measures. For example, our analysis with heart failure shows that value-based purchasing in the U.S. Medicare may have had no effect on change in performance. The fact may be explained by high-costs of hospitals with high quality measures for heart failure, while low-cost hospitals demonstrate the highest quality scores in cases pneumonia (Chen et al. (2010)). The finding with the U.S. data corresponds to the effect of payment-by-result in the U.K. where quality measures for heart failure were unaffected by the reform (Campbell et al. (2009)).

Finally, it should be noted that the major purpose of performance-based reimbursement reforms in the U.S. and Japan was the establishment of nationwide standards of health care and collection of nationwide databases (Werner and Dudley (2012)).[33] Indeed, the data for both countries indicate that skewness in the distribution of target indicators vanishes only after about a decade, and requires standardization criteria and performance monitoring.

8 Conclusion

The paper demonstrates differential effects of incentives regulations in health care, when the yardstick competition model is combined with performance-based reimbursement, related to benchmark values of quality measures or length-of-stay. We propose a theoretical model, forecasting the adverse effects for hospitals with the best target indicators.

The predictions of the model are estimated with a quantile regression approach, which accounts for several types of heterogeneous effects. To the best of our knowledge, the paper is the first extension of quantile regression methodology for dynamic panel data models with endogeneity. Our empirical analysis uses the U.S. and Japanese nationwide hospital/diagnosis-group level administrative panel data on a recent changeover to performance-based remuneration (in the U.S., Hospital Compare data for 4048 hospitals in 2008-2013, and in Japan Ministry of Health, Labor and Welfare data for 1849 hospitals in 2005-2013). The analysis with Japanese data focuses on the 2012/2013 natural experiment of changing benchmark values in the price schedule.

The results show persuasive evidence of the unwanted effects of incentives regulation for the best-performing hospitals. Each quality measure for patient experience of care in U.S. hospitals significantly decreases in the top percentiles of hospitals. Similarly, length of stay of Japanese hospitals significantly increases for hospitals in percentiles with the lowest nationwide length of stay. A natural experiment with a step toward best-practice rate-setting decreases the extent of performance deterioration for the benchmark hospitals.

[32] Despite the limitations of per diem rates, a changeover to full PPS would be a premature measure in Japan. So Japan sustains the per diem character of its payment system, renaming it as a "diagnosis-procedure combination/per diem payment system" – DPC/PDPS (MHLW 2011b).

[33] So the current percentage of hospital funds at risk in U.S. hospitals (1-2%) is negligible.

Appendix A Price-setting in the U.S. value-based purchasing

Table 6: **Quality measures for VBP in 2013 and benchmark scores**

Measure	Definition	Threshold	Benchmark	Floor
HCAHPS				
Clean-hsp-ap	Room was always clean	62.80	77.64	36.88
Comp-1-ap	Nurses always communicated well	75.18	84.70	38.98
Comp-2-ap	Doctors always communicated well	79.42	88.95	51.51
Comp-3-ap	Patients always received help as soon as they wanted	61.82	77.69	30.25
Comp-4-ap	Pain was always well controlled	68.75	77.90	34.76
Comp-5-ap	Staff always gave explanation about medicines	59.28	70.42	29.27
Comp-6-yp	Yes, staff did give patients discharge information	81.93	89.09	50.47
Hsp-rating-910	Patients who gave hospital a rating of 9 or 10 (high)	66.02	82.52	29.32
Quiet-hsp-ap	Hospital always quiet at night	62.80	77.64	36.88
Clinical process of care				
AMI-7a	Fibrinolytic therapy received within 30 minutes of hospital arrival	65.48	91.91	
AMI-8a	Primary percutaneous coronary intervention received within 90 minutes of Hospital Arrival	91.86	100.00	
HF-1	Discharge instructions	90.77	100.00	
PN-3b	Blood cultures performed in the emergency department prior to initial antibiotic received in hospital	96.43	100.00	
PN-6	Initial antibiotic selection for CAP in immunocompetent patient	92.77	99.58	
SCIP-Card2	Surgery patients on beta-blocker therapy prior to arrival who received a beta-blocker during the perioperative period	97.35	99.58	
SCIP-Inf1	Prophylactic antibiotic received within 1 hour prior to surgical incision	97.66	100.00	
SCIP-Inf2	Prophylactic antibiotic selection for surgical patients	95.07	99.68	
SCIP-Inf3	Prophylactic antibiotics discontinued within 24 hours after surgery end time	94.28	99.63	
SCIP-Inf4	Cardiac surgery patients with controlled 6 A.M. postoperative blood glucose	95.00	100.00	
SCIP-VTE1	Surgery patients with recommended venous thromboembolism prophylaxis ordered	93.07	99.85	
SCIP-VTE2	Surgery patients who received appropriate venous thromboembolism prophylaxis within 24 Hours prior to surgery to 24 hours after surgery	93.99	100.00	

Note: Threshold is the percentage point score at 50th percentile, benchmark is score at the mean of top decile, floor is the minimum score based on survey of 3211 hospitals in the baseline period (Jul 2009-Mar 2010). Source: FY 2013 final rule. Federal Register, Vol.76, No.88, May 6, 2011, Tables 4 and 9.

Appendix B Health care system in Japan

Since 1961 Japan has had a mandatory and universal social health insurance, which has resulted in the expansion of health care utilization and improvement of the population's health status (Kondo and Shigeoka (2013); Ikegami et al. (2011)). Enrollment in one of the mutually exclusive health insurance plans is obligatory and depends on an enrollee's age and status at the labor market. The following health insurance plans exist in Japan: 1) national health insurance, which is municipality-managed insurance for the self-employed, retirees and their dependents; 2) government-managed insurance for employees of small firms and their dependents; 3) company-managed insurance associations created by firms with over 300 workers for their employees and employees' dependents; 4) benefit schemes set up by mutual aid associations.

Japanese health insurance is based on free access. The users of any health insurance plan can choose any health care institution, regardless of its location or type (e.g., private/public, hospital, clinic or ambulatory division of hospital). There are no gatekeepers, and payments for seeking the services of a large facility without referral are negligible (Ikegami and Campbell (1995)). The medical services and drugs covered by health insurance and their costs are listed in the national fee schedule, which is revised biennially by the Ministry of Health, Labor, and Welfare (MHLW) following the recommendations of its advisory committee – the Central Social Insurance Medical Council (Ministry of Health, Labor and Welfare (2013); Ikegami (2006); Campbell and Ikegami (1998); Bhattacharya et al. (1996)). The schedule includes four parts: medical services; dental services; drugs and materials; and prospective fees for inpatient care (since 2003). The predecessor of the current fee schedule is the schedule developed for office based physicians upon the introduction of health insurance for manual workers in 1927 (Campbell and Ikegami (1998)). Note that the 1961 adoption of the universal health insurance retained the coexistence of the old fee schedule, favoring private practitioners and exploited by clinics and small hospitals, and the new schedule, supporting specialized care and used by hospitals (National Institute of Population and Social Security Research (2005); Campbell and Ikegami (1998); Ikegami (1991)). Additionally, an establishment of a separate health insurance for the elderly in 1982 led to an adoption of a special fee schedule for financing the treatment of this age cohort (National Institute of Population and Social Security Research (2005); Ikegami (1991)). The three schedules are set by the MHLW, and the differences between the three schedules are minor (Ikegami (1991)). The old and new schedules were combined in 1994, and therefore, currently, a unified national fee schedule applies to all health care providers (Ikegami et al. (2011)).

Cost containment entered the agenda of Japanese health care policy makers in the 1970s, when the rate of health care expenditure growth started to exceed the rate of growth of GDP (Fujii and Reich (1988)) and health care system became highly subsidized. The factors causing soaring health care costs are population aging and the spread of new medical technologies in the environment of physician-induced demand under fee-for-service reimbursement. By the early 2000s the effectiveness of raising coinsurance rates and lowering of fees in the unified fee schedule had been exhausted as means of containing health care costs (Ikegami (2009)). Consequently, the Ministry of Health, Labor, and Welfare (MHLW) decided to introduce an inpatient prospective payment system for acute care hospitals to create incentives for cost containment.

Appendix C Diagnosis-procedure combinations in Japan

Table C.1: **Major Diagnostic Categories in Japan**

MDC	Definition	Notes
01	Nervous system	
02	Eye	
03	Ear, nose, mouth and throat	
04	Respiratory system	
05	Circulatory system	
06	Alimentary, liver, biliary-tree, and pancreas	MDC06 and MDC07 in ICD-10
07	Musculoskeletal system	
08	Skin and subcutaneous tissue	A part of MDC09 in ICD-10
09	Breast	A part of MDC09 in ICD-10
10	Endocrine, nutritional and metabolic system	
11	Kidney and urinary tract and male reproductive system	MDC11 and MDC12 in ICD-10
12	Female reproductive system and puerperal diseases, abnormal pregnancy, abnormal labor	MDC13 and MDC14 in ICD-10
13	Blood and blood forming organs and immunological disorders	MDC16 and MDC17 in ICD-10
14	Newborn and other neonates, congenital anomalies	
15	Pediatric diseases	
16	Trauma, burns, poison	Since 2008
17	Mental diseases and disorders	Since 2008
18	Miscellaneous	

Notes: English equivalents of MDCs in Ministry of Health, Labor and Welfare (2014a) are adopted from Hayashida et al. (2009), Kuwabara et al. (2008) and Ishikawa et al. (2005)

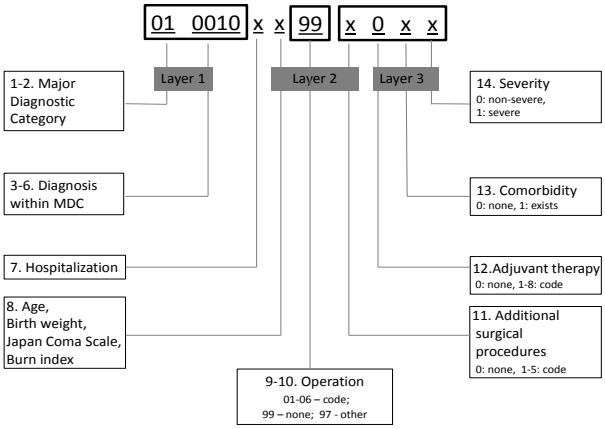

Figure C.1: **Code of Japanese DPC 010010xx99x0xx:**
"Nervous system diseases (MDC01), brain tumor (0010), no operation, no adjuvant therapy"
Source:Ministry of Health, Labor and Welfare (2014a)

Appendix D Estimating dynamic panel data models

Table D.1: **Threshold and the values of long-term mean for quality measures in dynamic panel data quantile regression ($\tau = 0.5$) and OLS models**

	HCAHPS measures									Clinical process of care measures				
	Comp-1ap	Comp-2ap	Comp-3ap	Comp-4ap	Comp-5ap	Clean-hsp-ap	Quiet-hsp-ap	Hrecomddy	Hsp-rating910	AMI-8a	HF-1	SCIP-Inf1	SCIP-Inf3	SCIP-Inf4	SCIP-VTE2
Threshold	75.18	79.42	61.82	68.75	59.28	62.80	62.80	n.a.	66.02	91.86	90.77	97.66	94.28	95.00	93.99
Quantile regression															
$\hat{\mu}_r$	78.271***	84.586***	66.803***	70.734***	64.209***	69.679***	62.508***	66.205***	65.833***	1488.2	70.678***	629.95	69.297***	112.46**	72.84***
	(4.455)	(5.051)	(4.047)	(4.349)	(5.557)	(3.86)	(4.141)	(4.14)	(3.764)	(36768)	(9.446)	(8850.2)	(12.168)	(49.818)	(13.415)
$\hat{\mu}_n$	79.059***	85.193***	68.217***	71.161***	66.456***	70.222***	63.953***	66.555***	66.291***	1545.1	67.2***	646.96	67.866***	109.36**	73.361***
	(4.493)	(5.058)	(4.104)	(4.371)	(5.638)	(3.876)	(4.187)	(4.139)	(3.774)	(38128)	(9.041)	(9116.6)	(11.832)	(48.374)	(13.438)
OLS model															
$\hat{\mu}_r$	78.05***	84.51***	66.98***	71.10***	62.30***	66.72***	58.33***	65.58***	65.25***	52.78***	19.64***	53.02***	22.46***	75.39***	90.54***
	(4.389)	(4.993)	(3.825)	(3.472)	(3.243)	(4.830)	(3.747)	(4.440)	(3.968)	(17.32)	(8.684)	(9.496)	(9.307)	(14.19)	(9.584)
$\hat{\mu}_n$	78.71***	86.12***	68.45***	71.90***	63.07***	65.76***	59.44***	65.25***	64.93***	45.83***	-12.20	57.13***	-6.201	60.20***	103.2***
	(4.789)	(6.789)	(4.280)	(5.234)	(5.576)	(5.130)	(4.252)	(4.825)	(4.342)	(24.51)	(13.60)	(13.87)	(14.64)	(18.58)	(15.13)

Note: Robust standard errors, estimated using delta method, in parentheses. $\hat{\mu}_r$ and $\hat{\mu}_n$ denote the estimated value of long-term mean for value-based purchasing participant and non-participant hospitals, repectively. Bound values of lagged dependent variable in quantile regressions for AMI-8a and SCIP-Inf1 may result in insignificant values of the long-term means.

Table D.2: **Estimation of dynamic panel data OLS model for quality measures**

	HCAHPS measures									**Clinical process of care measures**					
	Comp-1ap	Comp-2ap	Comp-3ap	Comp-4ap	Comp-5ap	Clean-hsp-ap	Quiet-hsp-ap	Hrecomddy	Hsp-rating910	AMI-8a	HF-1	SCIP-Inf1	SCIP-Inf3	SCIP-Inf4	SCIP-VTE2
L(score)	0.160***	0.134***	0.224***	0.178***	0.055	0.212***	0.270***	0.246***	0.183***	0.123***	0.285***	0.162***	0.221***	0.081	0.241***
	(0.024)	(0.033)	(0.022)	(0.028)	(0.043)	(0.020)	(0.023)	(0.026)	(0.024)	(0.035)	(0.025)	(0.026)	(0.034)	(0.054)	(0.032)
VBP·L(score)	-0.226***	-0.248***	-0.184***	-0.318***	-0.277***	-0.220***	-0.137***	-0.113***	-0.147***	-0.411***	-0.452***	-0.382***	-0.443***	-0.355***	-0.387***
	(0.013)	(0.014)	(0.012)	(0.018)	(0.026)	(0.014)	(0.008)	(0.009)	(0.011)	(0.039)	(0.028)	(0.029)	(0.038)	(0.060)	(0.035)
VBP	17.089***	19.558***	11.172***	21.918***	16.508***	15.428***	7.164***	7.671***	9.837***	27.758***	31.647***	16.787***	32.260***	40.729***	25.473***
	(1.034)	(1.150)	(0.826)	(1.274)	(1.769)	(0.925)	(0.575)	(0.701)	(0.790)	(7.437)	(3.337)	(3.642)	(4.467)	(9.809)	(3.914)
public	0.255	-0.457	0.538	-0.122	0.371	0.648	0.002	0.010	-0.368	-11.953***	0.462	5.851*	2.829	-4.057	4.195
	(0.277)	(0.312)	(0.448)	(0.397)	(0.428)	(0.396)	(0.412)	(0.400)	(0.405)	(4.092)	(2.820)	(3.372)	(2.005)	(2.956)	(2.709)
emergency	0.139	-0.218	-0.086	0.024	-0.009	0.359	0.006	-0.571*	0.077	-5.755	0.498	3.599	1.216	0.390	1.545
	(0.216)	(0.198)	(0.317)	(0.256)	(0.336)	(0.289)	(0.311)	(0.342)	(0.330)	(4.236)	(2.037)	(2.880)	(1.550)	(4.464)	(2.101)
urban	-1.697***	-1.715**	-1.329	-0.821	-1.378	1.025	1.369	5.617***	2.805***	18.817**	2.267	-4.487	19.274**	6.548	-10.709*
	(0.631)	(0.803)	(0.824)	(0.663)	(0.926)	(0.991)	(1.012)	(0.915)	(0.875)	(9.440)	(4.336)	(5.568)	(7.665)	(5.805)	(6.464)
teaching	0.189	-0.278	-0.261	0.206	-0.392	0.425	-0.781**	0.190	0.475	15.556***	8.678***	11.067***	11.785***	10.563**	-4.598
	(0.271)	(0.268)	(0.427)	(0.378)	(0.373)	(0.369)	(0.367)	(0.426)	(0.383)	(5.145)	(3.151)	(3.786)	(3.668)	(4.147)	(3.079)
casemix	2.117***	0.388	3.802***	1.823***	3.100***	4.992***	4.315***	5.125***	5.526***	-8.761	16.830***	7.442	10.274**	-1.775	-13.947***
	(0.519)	(0.497)	(0.899)	(0.621)	(0.820)	(1.718)	(0.853)	(0.728)	(0.734)	(9.405)	(5.074)	(5.193)	(4.761)	(7.141)	(4.752)
beds	-0.012***	-0.011***	-0.020***	-0.011***	-0.015***	-0.008***	-0.023***	-0.013***	-0.015***	0.073***	0.114***	0.011	0.169***	0.065***	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.023)	(0.016)	(0.013)	(0.018)	(0.015)	(0.013)
medicare_share	-5.738***	-5.308***	-10.736***	-5.201***	-8.470***	-7.886***	-8.761***	-11.676***	-12.428***	-2.091	26.163***	13.701	9.188	-10.761	28.453***
	(1.003)	(0.841)	(1.519)	(1.224)	(1.643)	(1.470)	(1.448)	(1.408)	(1.429)	(16.305)	(9.534)	(10.577)	(8.405)	(12.384)	(8.399)
year2010	0.958***	0.298***	0.486***	0.323***	1.195***	0.469***	0.634***	0.026	0.881***	-6.466***	-3.418***	-9.368***	-3.844***	-1.742**	-0.185
	(0.076)	(0.060)	(0.094)	(0.076)	(0.092)	(0.102)	(0.096)	(0.089)	(0.109)	(1.182)	(0.664)	(0.791)	(0.635)	(0.801)	(0.603)
year2011	2.101***	0.976***	2.540***	1.103***	2.992***	0.994***	2.360***	1.080***	2.055***	-5.321	0.719	0.359	-1.192	-11.941*	6.434**
	(0.182)	(0.181)	(0.303)	(0.264)	(0.389)	(0.385)	(0.283)	(0.272)	(0.280)	(6.699)	(2.400)	(2.956)	(2.891)	(7.247)	(2.649)
year2012	2.591***	1.098***	2.523***	1.088***	3.637***	1.184***	2.378***	0.515**	2.189***	-14.590**	-4.223*	-13.528***	-9.147***	-17.020**	-2.990
	(0.182)	(0.175)	(0.302)	(0.260)	(0.416)	(0.347)	(0.283)	(0.256)	(0.268)	(6.645)	(2.375)	(2.854)	(2.824)	(7.197)	(2.620)
Constant	66.110***	74.555***	53.130***	59.077***	59.612***	51.843***	43.374***	49.178***	53.037***	40.198*	-8.723	47.884***	-4.827	55.342***	78.329***
	(2.253)	(3.159)	(2.129)	(2.468)	(2.868)	(3.053)	(2.145)	(2.251)	(2.279)	(21.284)	(9.770)	(11.304)	(11.433)	(15.806)	(10.412)
Observations	12,701	12,701	12,700	12,694	12,695	12,701	12,700	12,700	12,699	5,213	11,974	11,731	11,693	4,464	11,729
Hospitals	3,290	3,290	3,290	3,289	3,289	3,290	3,290	3,289	3,289	1,437	3,117	3,061	3,057	1,159	3,084
$\hat{\alpha}_1 + \hat{\alpha}_2$	-0.066***	-0.114***	0.040***	-0.139***	-0.222***	-0.008***	0.134***	0.133***	0.037	-0.288***	-0.167***	-0.220***	-0.221***	-0.274***	-0.146***
	(0.028)	(0.032)	(0.024)	(0.025)	(0.030)	(0.026)	(0.024)	(0.027)	(0.026)	(0.025)	(0.022)	(0.018)	(0.024)	(0.040)	(0.022)

Note: Robust standard errors (estimated for $\hat{\alpha}_1 + \hat{\alpha}_2$ using delta method) in parentheses. L(score) denotes the lagged score for each quality measure. Samples are slightly lower than in quantile regressions, owing to the presence of VBP·L(score) and the need to identify its lag in the instrumental variable estimations. *** p<0.01, ** p<0.05, * p<0.1

Table D.3: **Effect of VBP on quality measures at group means (OLS model)**

	HCAHPS measures									**Clinical process of care measures**					
	Comp-1ap	Comp-2ap	Comp-3ap	Comp-4ap	Comp-5ap	Clean-hsp-ap	Quiet-hsp-ap	Hrecomddy	Hsp-rating910	AMI-8a	HF-1	SCIP-Inf1	SCIP-Inf3	SCIP-Inf4	SCIP-VTE2
decile1	1.619***	1.358***	1.478***	1.833***	1.395***	1.836***	1.106***	1.516***	1.818***	5.365	8.542***	-5.315*	3.333	15.27**	1.766
	(0.225)	(0.200)	(0.319)	(0.286)	(0.433)	(0.314)	(0.296)	(0.310)	(0.308)	(6.812)	(2.600)	(2.926)	(3.035)	(7.624)	(2.736)
decile2	0.597***	0.574***	0.433***	0.603***	0.352***	1.000***	0.373	0.787***	0.863***	n.a.	12.67***	n.a.	-1.662	9.427	0.844
	(0.196)	(0.183)	(0.292)	(0.260)	(0.365)	(0.313)	(0.278)	(0.283)	(0.276)	n.a.	(2.690)	n.a.	(2.931)	(7.391)	(2.715)
decile3	0.174	0.264	0.0939	0.301	0.164	0.611*	0.010	0.405	0.387	n.a.	-1.117	n.a.	-6.596***	8.233	-7.109***
	(0.187)	(0.179)	(0.286)	(0.255)	(0.355)	(0.315)	(0.271)	(0.273)	(0.266)	n.a.	(2.484)	n.a.	(2.887)	(7.359)	(2.633)
decile4	-0.055	-0.029	-0.250	0.153	-0.263	0.262	-0.322	0.140	0.118	-3.192	-5.678**	-14.04***	-7.518***	8.439	-9.477***
	(0.184)	(0.177)	(0.282)	(0.254)	(0.333)	(0.319)	(0.267)	(0.268)	(0.262)	(6.735)	(2.480)	(2.867)	(2.886)	(7.364)	(2.646)
decile5	-0.342*	-0.267	-0.406	-0.208	-0.533*	-0.058	-0.719***	-0.178	-0.122	-3.143	-6.461***	-16.47***	-8.082***	8.175	-10.07***
	(0.181)	(0.176)	(0.280)	(0.251)	(0.321)	(0.324)	(0.263)	(0.264)	(0.259)	(6.736)	(2.483)	(2.877)	(2.886)	(7.358)	(2.652)
decile6	-0.504***	-0.508***	-0.791***	-0.400	-0.587*	-0.323	-0.996***	-0.430	-0.370	-5.424	-7.771***	-16.08***	-8.477***	8.666	-9.829***
	(0.180)	(0.177)	(0.278)	(0.250)	(0.319)	(0.329)	(0.262)	(0.263)	(0.258)	(6.731)	(2.489)	(2.874)	(2.887)	(7.370)	(2.649)
decile7	-0.715***	-0.856***	-1.087***	-0.543**	-0.985***	-0.687**	-1.324***	-0.715***	-0.658***	-3.185	-6.909***	-14.000***	n.a.	8.745	n.a.
	(0.179)	(0.179)	(0.278)	(0.249)	(0.305)	(0.336)	(0.263)	(0.263)	(0.258)	(6.735)	(2.485)	(2.867)	n.a.	(7.372)	n.a.
decile8	-0.955***	-1.100***	-1.288***	-0.876***	-1.225***	-1.051***	-1.731***	-0.989***	-0.910***	-1.987	-7.092***	n.a.	-6.370**	n.a.	-7.568***
	(0.179)	(0.182)	(0.279)	(0.249)	(0.298)	(0.345)	(0.265)	(0.264)	(0.260)	(6.740)	(2.485)	n.a.	(2.888)	n.a.	(2.634)
decile9	-1.238***	-1.435***	-1.876***	-1.285***	-1.458***	-1.525***	-2.192***	-1.317***	-1.260***	-1.176	-5.669**	n.a.	n.a.	9.927	n.a.
	(0.181)	(0.188)	(0.284)	(0.251)	(0.293)	(0.359)	(0.271)	(0.269)	(0.264)	(6.745)	(2.480)	n.a.	n.a.	(7.406)	n.a.
decile10	-2.032***	-2.177***	-3.028***	-2.387***	-2.581***	-2.514***	-3.224***	-2.000***	-2.247***	-4.117	-3.500	n.a.	n.a.	n.a.	n.a.
	(0.192)	(0.207)	(0.309)	(0.266)	(0.292)	(0.392)	(0.294)	(0.285)	(0.287)	(6.733)	(2.478)	n.a.	n.a.	n.a.	n.a.
p96-100	-2.725***	-2.485***	-4.173***	-3.080***	-3.554***	-3.180***	-4.214***	-2.392***	-2.860***						
	(0.210)	(0.216)	(0.347)	(0.281)	(0.321)	(0.419)	(0.326)	(0.299)	(0.309)						

Note: Robust standard errors, estimated using delta method, in parentheses. p96-100 denotes percentiles 96-100. N.a.=non-available (hospitals from corresponding percentiles of quality measures did not participate in VBP). *** p<0.01, ** p<0.05, * p<0.1

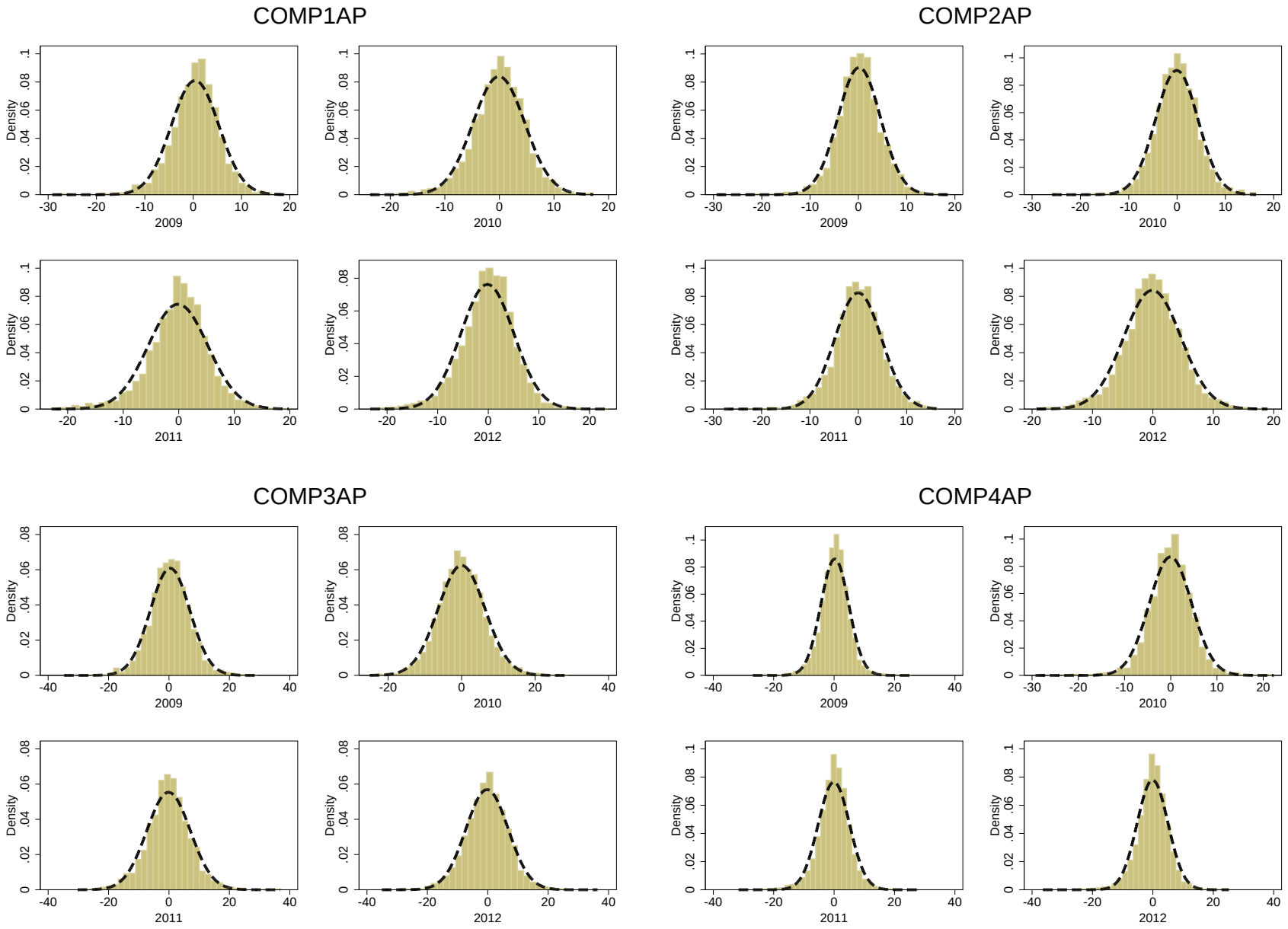

Figure D.1: Residuals in the OLS model for HSAHPS measures (dashed line is an attempt to fit normal pdf)

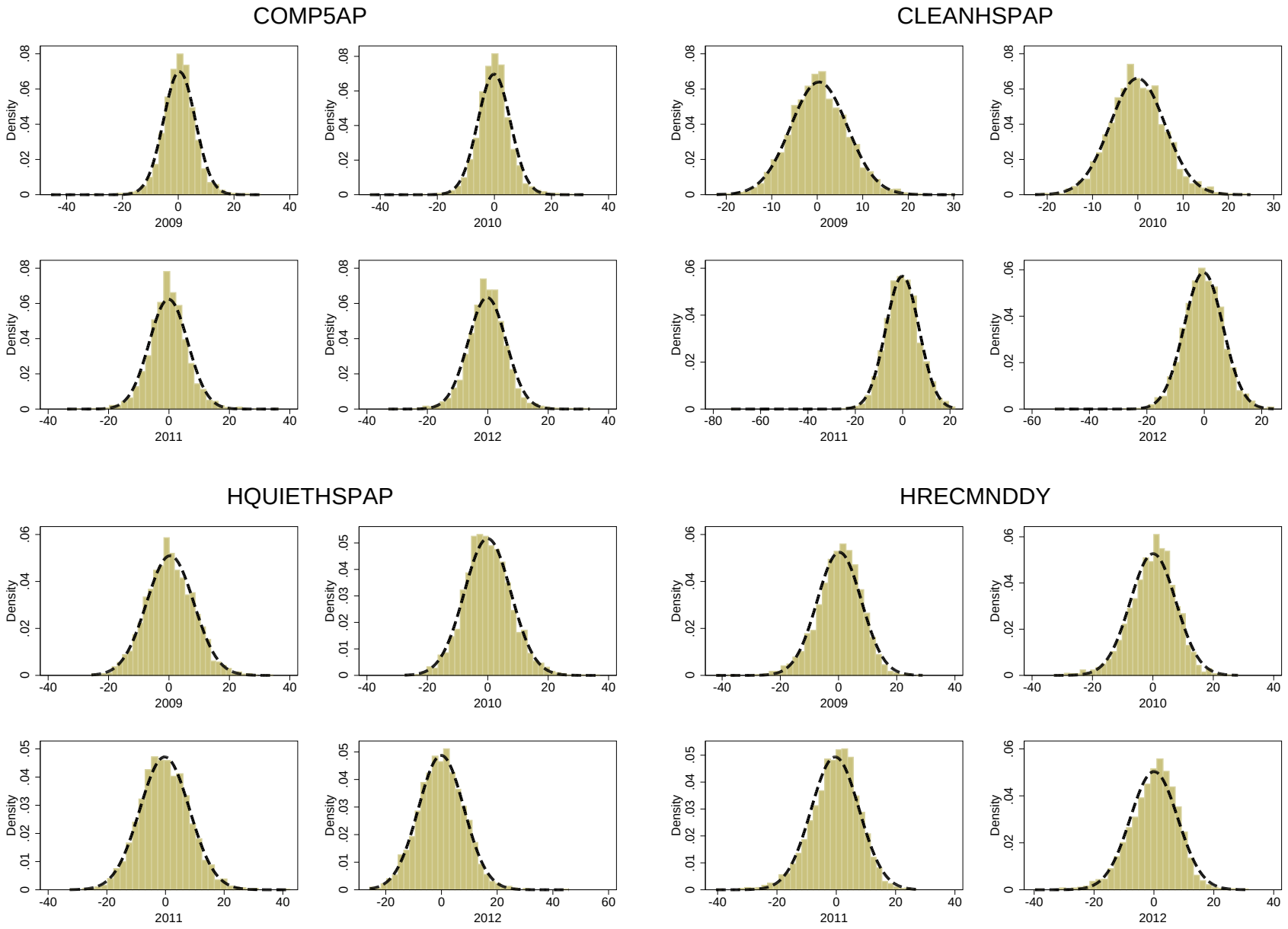

Figure D.2: Residuals in the OLS model for HSAHPS measures (dashed line is an attempt to fit normal pdf)

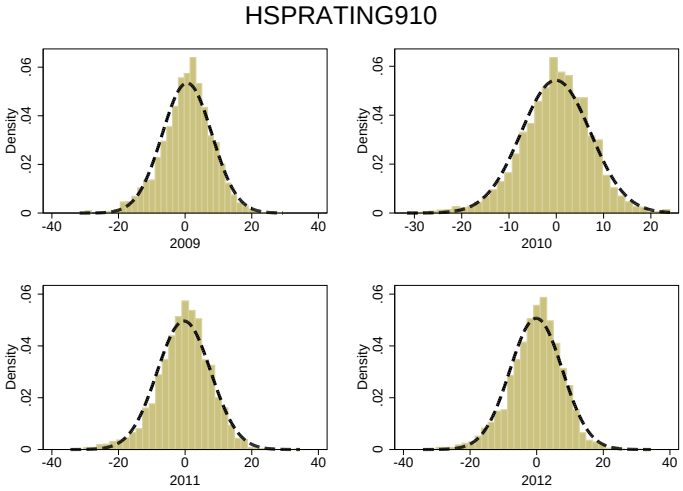

Figure D.3: Residuals in the OLS model for HSAHPS measures

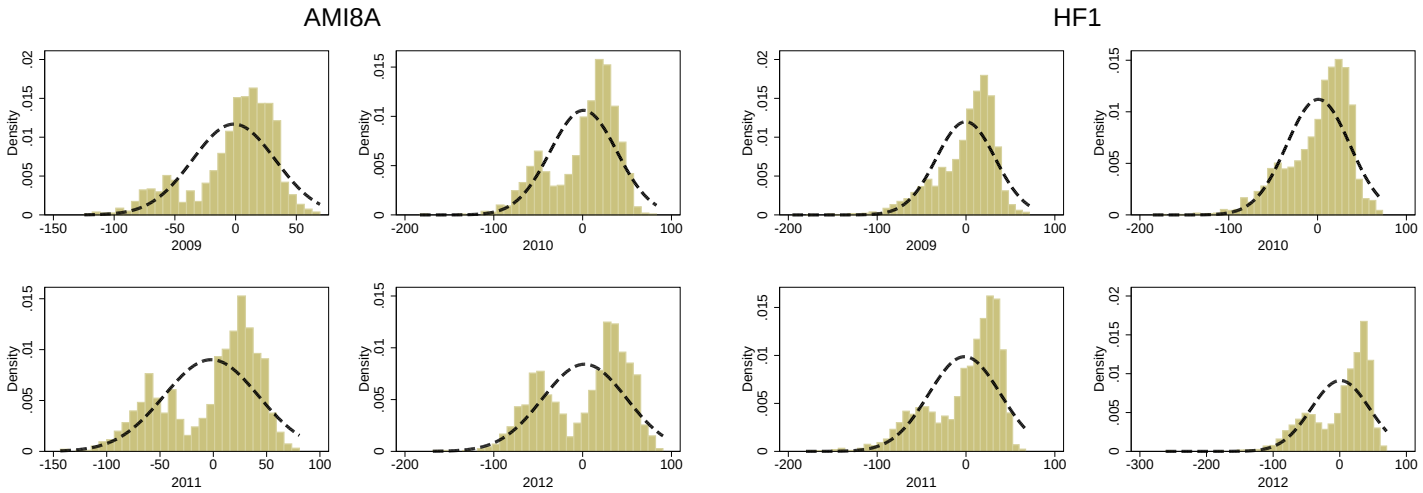

Figure D.4: Residuals in the OLS model for clinical process of care measures (dashed line is an attempt to fit normal pdf)

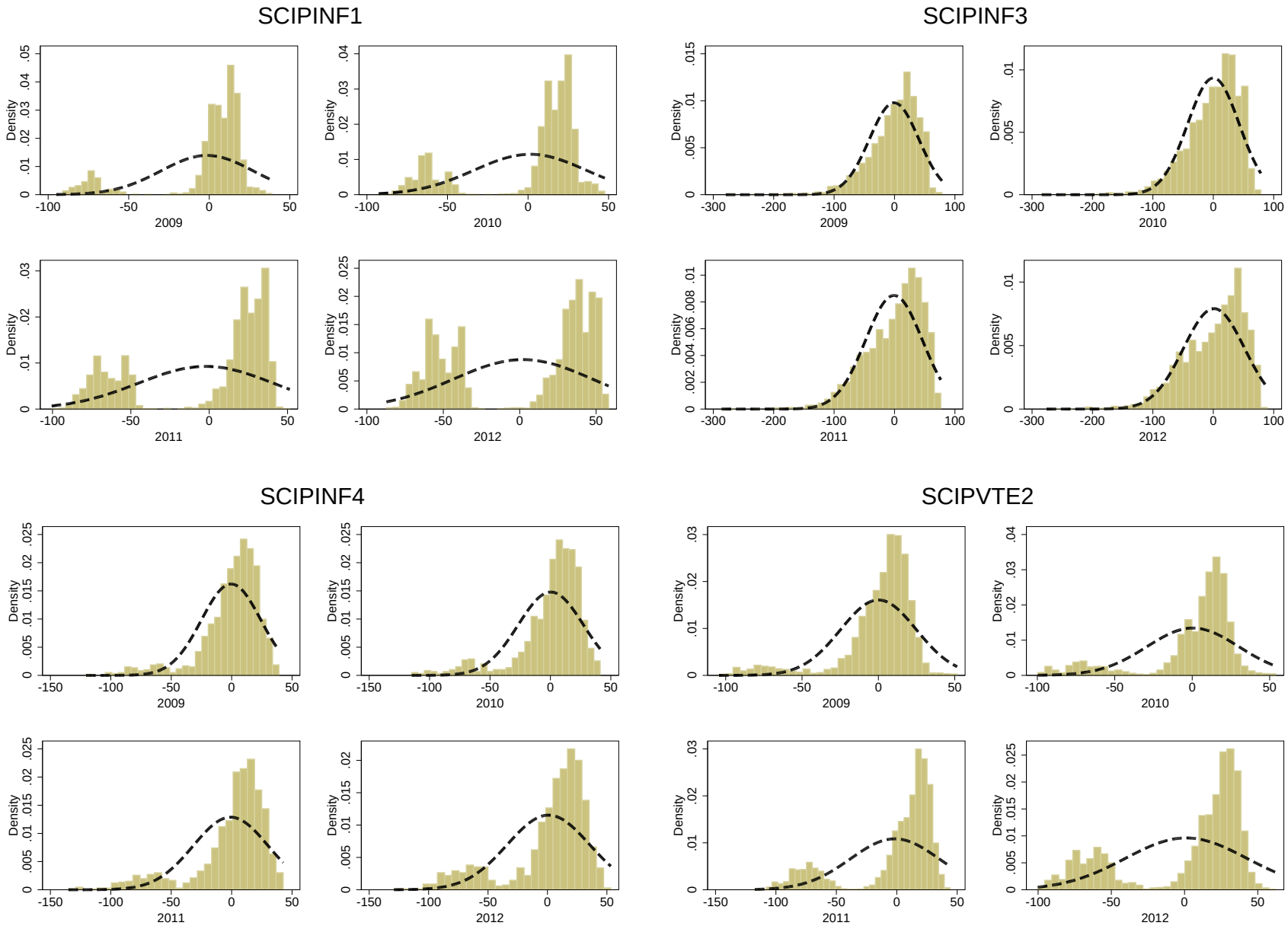

Figure D.5: Residuals in the OLS model for clinical process of care measures (dashed line is an attempt to fit normal pdf)

Table D.4: **Estimation of dynamic panel data OLS model for average length of stay (LOS) for each major diagnostic category (MDC)**

	MDC1	MDC2	MDC3	MDC4	MDC5	MDC6	MDC7	MDC8	MDC9	MDC10	MDC11	MDC12	MDC13	MDC14	MDC15	MDC16
L(los)	0.302***	0.732***	0.477***	0.718***	0.385***	0.539***	0.396***	-0.068	0.311***	0.072	0.317***	0.447***	0.400***	-0.063	0.023	0.258**
	(0.053)	(0.037)	(0.050)	(0.109)	(0.062)	(0.044)	(0.067)	(0.213)	(0.045)	(0.107)	(0.053)	(0.058)	(0.050)	(0.116)	(0.105)	(0.106)
L^2(los)	–	–	–	–	–	–	–	0.068	–	-0.072	–	–	–	-0.048	-0.039	–
	–	–	–	–	–	–	–	(0.211)	–	(0.088)	–	–	–	(0.195)	(0.109)	–
PPS·L(los)	-0.321***	-0.364***	-0.297***	-0.594***	-0.386***	-0.384***	-0.438***	-0.076	-0.458***	-0.259**	-0.339***	-0.321***	-0.459***	-0.098	-0.297***	-0.255**
	(0.052)	(0.038)	(0.061)	(0.093)	(0.059)	(0.042)	(0.067)	(0.217)	(0.048)	(0.107)	(0.054)	(0.061)	(0.050)	(0.124)	(0.105)	(0.102)
PPS·L^2(los)	–	–	–	–	–	–	–	-0.214	–	-0.055	–	–	–	0.110	-0.152	–
	–	–	–	–	–	–	–	(0.215)	–	(0.087)	–	–	–	(0.196)	(0.109)	–
PPS	6.689***	1.875***	2.183***	10.367***	5.436***	4.903***	8.003***	1.229	5.431***	3.311	4.185***	3.625***	11.023***	1.058	3.265**	3.657*
	(1.109)	(0.236)	(0.483)	(1.743)	(0.917)	(0.655)	(1.463)	(6.284)	(0.679)	(2.815)	(0.853)	(0.729)	(1.251)	(1.498)	(1.593)	(2.124)
share_DPClist	12.357***	7.586***	21.639***	-32.510***	-27.099***	0.930	-5.910	-23.741***	-1.575	-8.438***	-29.214***	8.411***	16.890***	9.584***	-0.605	3.572
	(3.312)	(0.989)	(3.507)	(2.524)	(3.044)	(1.485)	(3.866)	(3.015)	(3.829)	(3.152)	(3.067)	(1.969)	(6.442)	(3.594)	(1.972)	(3.523)
designated	-0.340**	-0.000	-0.249***	0.114	0.055	-0.042	-0.335**	-0.004	0.125	-0.128	0.234**	0.008	-0.334	0.011	0.010	-0.030
	(0.133)	(0.045)	(0.077)	(0.101)	(0.097)	(0.051)	(0.132)	(0.155)	(0.134)	(0.106)	(0.107)	(0.076)	(0.203)	(0.097)	(0.085)	(0.108)
quality	-0.137	-0.004	0.075	0.830*	-0.305	-0.881***	1.753***	-1.955	-0.817	-1.192	-0.109	-0.803***	4.250***	1.656**	-1.849**	-0.760
	(0.513)	(0.167)	(0.265)	(0.476)	(0.466)	(0.273)	(0.611)	(1.377)	(0.641)	(0.776)	(0.398)	(0.294)	(1.045)	(0.717)	(0.734)	(0.672)
year2010	0.743***	0.036	-0.003	0.172*	-0.143	-0.010	0.149	–	-0.343**	–	-0.323***	0.154	-0.119	–	–	–
	(0.136)	(0.041)	(0.068)	(0.098)	(0.114)	(0.060)	(0.140)	–	(0.137)	–	(0.110)	(0.114)	(0.218)	–	–	–
year2011	0.590***	-0.402***	-0.753***	0.998***	1.030***	-0.348***	0.369**	–	-0.985***	–	0.674***	-0.523***	-1.101***	–	–	0.345**
	(0.169)	(0.056)	(0.119)	(0.120)	(0.129)	(0.068)	(0.159)	–	(0.175)	–	(0.128)	(0.096)	(0.291)	–	–	(0.137)
year2012	-1.201***	-0.636***	-1.616***	0.718***	0.624***	-1.280***	-0.532***	0.804***	-1.386***	-0.779***	0.105	-1.129***	-3.571***	-1.021***	-0.412***	-0.893***
	(0.186)	(0.070)	(0.159)	(0.135)	(0.168)	(0.082)	(0.191)	(0.154)	(0.247)	(0.109)	(0.154)	(0.119)	(0.364)	(0.149)	(0.093)	(0.176)
year2013	-1.403***	-0.713***	-1.526***	1.092***	0.395**	-1.303***	-0.585***	0.317*	-1.583***	-1.133***	-0.119	-1.174***	-4.072***	-1.070***	-0.300***	-1.194***
	(0.210)	(0.081)	(0.167)	(0.144)	(0.176)	(0.108)	(0.220)	(0.169)	(0.256)	(0.127)	(0.176)	(0.154)	(0.380)	(0.162)	(0.092)	(0.191)
Constant	11.862***	-0.118	0.008	10.692***	14.780***	7.097***	12.704***	22.904***	8.775***	20.406***	16.388***	4.351***	8.654***	7.539***	9.822***	14.005***
	(1.229)	(0.291)	(0.847)	(2.242)	(1.194)	(0.734)	(1.825)	(5.999)	(1.254)	(3.142)	(1.014)	(0.840)	(1.894)	(1.637)	(1.669)	(2.427)
Observations	7,265	4,747	6,659	7,436	7,131	7,377	7,111	3,064	4,287	4,148	7,179	4,324	5,914	2,139	3,586	6,002
Hospitals	1,574	1,082	1,491	1,592	1,558	1,578	1,554	1,163	1,007	1,477	1,558	944	1,396	769	1,309	1,569
$\hat{\alpha}_1 + \hat{\alpha}_2$	-0.019	0.367***	0.180***	0.123***	-0.0003	0.155***	-0.042	-0.145***	-0.147***	-0.187***	-0.022	0.126***	-0.059***	-0.161***	-0.274***	0.004
	(0.025)	(0.029)	(0.046)	(0.029)	(0.023)	(0.022)	(0.026)	(0.030)	(0.025)	(0.028)	(0.024)	(0.032)	(0.019)	(0.048)	(0.032)	(0.023)

Note: Robust standard errors (estimated for $\hat{\alpha}_1 + \hat{\alpha}_2$ using delta method) in parentheses. L(los) and L^2(los) denote, respectively, the first and second lag of length of stay. Samples are slightly lower than in quantile regressions, owing to the presence of PPS·L(los) and the need to identify its lag in the instrumental variable estimations. Specifications with AR(2) use fewer annual dummies, since third and forth lags are necessary as instruments for quantile regressions. The first year in the differenced equation is a reference category. Fewer years are used in the AR(2) specification, so that the L^3(los) and L^4(los) were available for further estimates with quantile regression. *** p<0.01, ** p<0.05, * p<0.1

Table D.5: **Effect of inpatient prospective payment system on average length of stay for each major diagnostic category at group means in OLS model**

	MDC1	MDC2	MDC3	MDC4	MDC5	MDC6	MDC7	MDC8	MDC9	MDC10	MDC11	MDC12	MDC13	MDC14	MDC15	MDC16
$\hat{\delta}_1$ (new benchmark)	2.762***	0.66***	0.734***	4.239***	2.391***	1.375***	2.755***	-0.788	2.613***	1.17	1.302***	1.357***	4.248***	1.222	1.662	1.133
	(0.496)	(0.121)	(0.204)	(0.797)	(0.466)	(0.279)	(0.681)	(4.105)	(0.415)	(1.851)	(0.409)	(0.311)	(0.58)	(0.98)	(1.06)	(1.143)
$\hat{\delta}_1$ (counterfactual old benchmark)	2.118***	0.592***	0.612***	3.327***	1.974***	0.789***	1.99***	-1.588	2.255***	-0.168	0.929***	1.044***	3.523***	1.199	0.639	0.664
	(0.402)	(0.115)	(0.183)	(0.658)	(0.407)	(0.219)	(0.571)	(3.148)	(0.385)	(1.365)	(0.354)	(0.258)	(0.518)	(0.982)	(0.788)	(0.966)
$\hat{\delta}_2$	0.868***	-0.143**	-0.153	1.772***	1.066***	0.0517	0.521	-1.946	1.258***	-0.738	0.219	0.51***	0.755**	1.203	0.328	-0.167
	(0.245)	(0.071)	(0.106)	(0.427)	(0.284)	(0.149)	(0.371)	(2.745)	(0.309)	(1.191)	(0.255)	(0.176)	(0.35)	(1.113)	(0.716)	(0.667)
$\hat{\delta}_3$	-0.937***	-1.26***	-1.518***	-1.298***	-0.701***	-1.218***	-1.975***	-2.889	-0.673***	-2.244***	-1.264***	-0.539***	-3.713***	1.259	-0.681	-1.692***
	(0.241)	(0.122)	(0.306)	(0.181)	(0.163)	(0.0894)	(0.214)	(1.897)	(0.234)	(0.914)	(0.143)	(0.146)	(0.515)	(2.106)	(0.567)	(0.347)
$\hat{\delta}_4$	-3.954***	-3.111***	-4.914***	-6.129***	-3.94***	-3.143***	-5.996***	-4.365***	-4.878***	-4.371***	-3.948***	-2.478***	-10.8***	1.211	-2.493***	-4.062***
	(0.661)	(0.301)	(0.989)	(0.852)	(0.553)	(0.244)	(0.71)	(2.548)	(0.497)	(1.171)	(0.471)	(0.463)	(1.204)	(3.496)	(0.711)	(1.041)

Notes: For each major diagnostic category hospital-level length of stay is weighted by the number of patient cases in computing the empirical nationwide distribution. In case of new benchmark, $\hat{\delta}_k = \hat{\alpha}_2 \bar{y}_{i,t-1} + \hat{\alpha}_3$ in equation (21) for groups $k = 1, ..., 3$ of hospitals, who have lagged length of stay not exceeding the upper boundaries for period I, II and III, respectively; and group 4 of hospitals with lagged length of stay above the length of period III (mean values of lagged lengths of stay for each hospital are taken over 2007-2013). In the analysis with counterfactual old benchmark group I includes hospitals with lagged length of stay not exceeding the 25-th percentile. Robust standard errors, estimated using delta method, in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table D.6: **Threshold and the values of long-term mean for length of stay in dynamic panel data quantile regression ($\tau = 0.5$) and OLS models**

	020210xx97	030240xx99	030390xx99	030440xx01	060245xx97	060330xx02	060570xx99	080011xx99	11013xxx99	110200xx02	110200xx04	11022xxx99	120060xx01	120060xx02	120070xx01	120070xx02	120090xx97	120100xx01	120150xx99	120160xx99	120220xx01	120260xx01
Threshold	8	5	10	12	6	7	7	10	7	10	8	7	10	6	10	7	10	8	11	7	3	10
Quantile regression																						
$\hat{\mu}_r$	-0.807	5.454***	n.a.	16.15***	6.825***	8.29***	8.198***	18.224	17.824***	10.814***	6.652**	9.585	11.239***	4.993**	12.13***	9.403	12.3***	8.205*	15.939***	14.373***	2.774**	8.032***
	(1.871)	(0.532)	n.a.	(5.447)	(2.472)	(1.791)	(1.163)	(30.689)	(4.75)	(1.677)	(3.249)	(1643.9)	(1.81)	(2.19)	(3.151)	(10.48)	(3.682)	(4.72)	(4.318)	(5.201)	(1.374)	(2.567)
$\hat{\mu}_n$	-0.105	5.792***	n.a.	21.609**	6.827***	8.512***	8.401***	20.246	19.509***	10.877***	7.871*	10.546	11.319***	4.42	12.781***	4.671	14.047***	8.105*	17.152***	13.717***	3.076*	8.033***
	(2.462)	(0.587)	n.a.	(11.542)	(2.677)	(2.9)	(1.162)	(34.736)	(5.986)	(1.985)	(4.342)	(1908.5)	(1.965)	(2.899)	(3.605)	(7.774)	(3.88)	(4.909)	(5.816)	(4.901)	(1.794)	(2.978)
OLS model																						
$\hat{\mu}_r$	3.587	6.477***	15.68***	18.37***	6.476***	11.37***	7.364***	15.07***	16.43***	14.32***	6.205**	8.432***	14.06***	7.091***	14.37***	6.939***	14.91***	7.805***	13.02***	12.04***	3.137***	9.81***
	(3.354)	(0.674)	(3.668)	(4.136)	(1.775)	(1.055)	(1.522)	(1.16)	(3.849)	(1.642)	(2.945)	(1.693)	(1.248)	(1.358)	(1.617)	(1.212)	(2.082)	(1.345)	(2.787)	(2.394)	(0.647)	(0.895)
$\hat{\mu}_n$	-0.454	7.013***	17.29***	22.13*	7.281*	16.05***	7.831	20.18	20.88***	18.13**	-6.021	9.401***	17.5***	6.831**	16.93***	7.473	16.58***	7.326***	12.54**	10.69**	3.264***	8.814
	(10.28)	(1.711)	(5.853)	(14.7)	(4.779)	(5.853)	(4.731)	(19.35)	(6.593)	(8.816)	(9.439)	(2.323)	(5.449)	(3.667)	(6.233)	(13.85)	(5.805)	(2.014)	(7.546)	(6.106)	(0.862)	(11.97)
Obs	256	748	695	895	571	2209	712	3500	166	1814	278	519	2959	1035	1991	1594	1094	732	613	332	289	1041
Hospitals	90	345	275	245	229	678	310	1003	74	545	98	217	713	329	534	450	400	282	224	145	123	340

Note: For each diagnosis-procedure combination (DPC) hospital-level length of stay is weighted by the number of patient cases in computing the empirical nationwide distribution. The Table demonstrates the effect for 22 groups of 10-digit DPCs, for which the specification holds (out of 32 groups with enough cases and price data available). Robust standard errors, estimated using delta method, in parentheses. $\hat{\mu}_r$ and $\hat{\mu}_n$ denote the estimated value of long-term mean for PPS participant and non-participant hospitals, respectively. No convergence was achieved in quantile regression with DPC "030390xx99".

Table D.7: **List of analysed diagnosis-procedure combinations (DPCs) in each major diagnostic category (MDC)**

MDC	DPC	Obs	Hospitals	MDC	DPC	Obs	Hospitals	MDC	DPC	Obs	Hospitals	MDC	DPC	Obs	Hospitals	MDC	DPC	Obs	Hospitals
1	010010xx01	946	311	3	030400xx99	4259	1172	6	060050xx03	1590	470	7	070560xx99	1717	494	12	120150xx99	613	224
1	010010xx99	1722	513	3	030430xx97	217	96	6	060050xx97	3994	953	8	080011xx99	3500	1003	12	120160xx99	332	145
1	010020xx01	763	246	3	030440xx01	895	245	6	060050xx99	4029	1016	9	090010xx97	2309	764	12	120170xx01	386	142
1	010020xx99	119	68	4	040010xx99	207	93	6	060060xx97	1361	496	9	090010xx99	2069	596	12	120170xx99	2005	556
1	010030xx01	505	183	4	040040xx01	2305	559	6	060060xx99	842	353	10	100020xx01	1095	320	12	120180xx01	1495	422
1	010030xx03	286	114	4	040040xx97	1361	436	6	060090xx02	628	250	10	100130xx97	264	105	12	120180xx99	352	133
1	010030xx99	1056	349	4	040040xx99	4820	1133	6	060100xx02	6069	1409	10	100393xx99	914	421	12	120220xx01	289	123
1	010040xx01	420	169	4	040080xx97	2257	746	6	060100xx99	2960	917	11	110060xx99	672	283	12	120260xx01	1041	340
1	010040xx99	3451	850	4	040080xx99	7319	1573	6	060130xx02	1328	498	11	110070xx02	4040	958	13	130010xx97	1328	348
1	010050xx02	1112	395	4	040120xx99	1137	488	6	060130xx97	911	370	11	110070xx99	1253	442	13	130030xx97	1642	432
1	010060xx01	429	157	4	040130xx99	1318	482	6	060130xx99	5059	1311	11	110080xx01	1202	447	13	130030xx99	2309	576
1	010060xx02	343	187	4	040200xx01	1143	384	6	060140xx97	723	388	11	110080xx99	4410	1018	13	130060xx97	277	131
1	010060xx97	1430	500	4	040200xx99	1829	618	6	060140xx99	1932	691	11	11012xxx02	1023	380	13	130090xx97	1039	514
1	010060xx99	6642	1481	5	050030xx99	319	134	6	060150xx02	3413	922	11	11012xxx04	1947	479	14	140010xx97	750	229
1	010080xx99	1414	456	5	050050xx02	1288	377	6	060150xx99	1312	472	11	11012xxx97	200	104	14	140010xx99	3217	727
1	010160xx99	827	268	5	050050xx99	5122	1133	6	060160xx02	5613	1313	11	11012xxx99	489	205	16	160100xx02	1102	452
1	010200xx01	118	62	5	050070xx01	1237	340	6	060170xx02	1075	435	11	11013xxx97	128	70	16	160100xx97	1029	420
1	010200xx99	124	66	5	050070xx99	2611	763	6	060180xx99	451	145	11	11013xxx99	166	74	16	160100xx99	2829	876
1	010230xx99	3778	951	5	050080xx99	962	343	6	060190xx99	1706	732	11	110200xx02	1814	545	16	160200xx02	428	201
2	020110xx97	4643	1054	5	050163xx03	577	187	6	060210xx97	2334	785	11	110200xx04	278	98	16	160610xx01	427	183
2	020150xx97	422	119	5	050170xx03	1657	552	6	060210xx99	6354	1437	11	11022xxx99	519	217	16	160610xx97	197	110
2	020160xx97	1235	314	5	050170xx99	1109	400	6	060245xx97	571	229	11	110310xx97	231	134	16	160620xx01	1774	581
2	020180xx97	952	277	5	050180xx01	663	284	6	060290xx99	612	222	11	110310xx99	5410	1378	16	160690xx99	2695	959
2	020200xx97	1144	319	5	050210xx97	3591	926	6	060295xx99	1490	427	11	110420xx97	654	260	16	160700xx97	1056	475
2	020210xx97	256	90	6	060010xx99	2033	574	6	060300xx97	2270	634	12	120010xx01	879	280	16	160720xx01	531	265
2	020220xx97	853	244	6	060020xx01	2622	930	6	060300xx99	2213	718	12	120010xx97	191	91	16	160740xx97	1525	597
2	020230xx97	636	235	6	060020xx97	1620	571	6	060330xx02	2209	678	12	120010xx99	2359	575	16	160760xx97	2359	872
2	020240xx97	626	208	6	060020xx99	4467	1138	6	060335xx02	3307	981	12	120060xx01	2959	713	16	160780xx97	558	257
3	03001xxx01	1198	331	6	060035xx01	4485	1107	6	060335xx99	1940	720	12	120060xx02	1035	329	16	160800xx01	2307	864
3	03001xxx97	810	229	6	060035xx03	663	260	6	060340xx03	4264	1123	12	120070xx01	1991	534	16	160800xx97	367	160
3	03001xxx99	1219	312	6	060035xx97	1561	561	6	060340xx99	2481	858	12	120070xx02	1594	450	16	160800xx99	235	137
3	030150xx97	788	269	6	060035xx99	4310	1152	6	060350xx99	1830	679	12	120090xx97	1094	400	16	160850xx01	283	139
3	030240xx99	748	345	6	060040xx01	3021	839	6	060570xx99	712	310	12	120100xx01	732	282	16	160850xx97	978	434
3	030300xx01	160	72	6	060040xx97	1154	432	7	070230xx01	2161	613	12	120110xx99	179	91	16	160980xx99	270	175
3	030390xx99	695	275	6	060040xx99	3320	952	7	070470xx99	1084	307	12	120130xx97	451	168	16	161060xx99	382	189

Note: The Table lists 10-digit codes for diagnosis-procedure combinations, analyzed in each major diagnostic category. Data for MDC15 at the DPC level are unavailable.

Appendix E Treatment effects

E.1 Value-based purchasing

Regarding value-based purchasing in the U.S., hospital's revenue in terms of the share of its Medicare's budget is γ_i and its maximal value is $(s-1)\cdot\alpha+1$. In our empirical analysis we use the treatment variable, which equals hospital's monetary gain (G) of quality improvement and demonstrates the returns-to-quality after the introduction of the reform. This way the policy environment implies that the effects of the reform vary based on the distance to incentive changes.

Case 1. For hospitals that joined the reform $G_i = \big((s-1)\cdot\alpha-(\gamma_i-1)\big)\cdot M_i$, where M_i is hospital's Medicare's budget. Note that $G(\cdot)$ is a linear monotonously decreasing function of γ_i and equals zero for hospitals with the maximal potential values of γ_i (i.e. with $tps_i = 100$).

Case 2. Hospitals that decided to postpone joining the reform receive their full budget M_i in a given year (essentially, their $\gamma_i = 1$). So $G_i = (s-1)\cdot\alpha\cdot M_i$.

E.2 Length-of-stay performance

Using similar approach for evaluating the returns-to-performance with the Japanese per diem payment prospective payment system, we construct the treatment variable as follows. The MHLW's price-setting rules described in Section 3.2 allow computing the values of per diem payments p^I, p^{II} and p^{III} for a standard DPC in terms of $\bar{p}$: $p^I = 1.15\bar{p}$, $p^{II} = c\bar{p}$, $p^{III} = 0.85c\bar{p}$, where $l_{0.25}$ and $\bar{l}$ are respectively, the 25-percentile and mean of the nationwide length of stay for a given DPC, and $c = 1-0.15/(\bar{l}-l_{0.25})$.[34]

Let l_i be the average length of stay for a given standard DPC at a hospital i. Then, if a hospital has b_i beds, the number of patients that can be treated within d days is $N_i = db_i/l_i$. (For simplicity, we assume that the demand for hospital care is unconstrained and hospital may sustain 100-percent rate of bed occupancy. Alternatively, we analytically compute $N(l_i)$ and condition bed occupancy rate not to exceed unity: $\dfrac{N(l_i)l_i}{db_i} \leq 1$). Since our data has the same value of d for all hospitals (days in one fiscal year), we set d equal to unity.

Below we define treatment variable as the monetary gain (G_i) of hospital's decreasing its length of stay l_i till a minimal potential value l_{min}. Empirically l_{min} can be derived as nationwide minimal length-of-stay at the "best-practice" hospital or as an arbitrary chosen bottom percentile point. Cases 1–3 refer to hospitals that have joined the new per diem payment system. Case 4 describers hospitals that are financed according to the old fee-for-service system.

Case 1. $l_i \leq l_{0.25}$

A hospital receives the highest per diem rate p^I for each of its b_i/l_i patients. Per patient revenue is $p^I l_i = 1.15pl_i$, so the total revenue R_i is $1.15pb_i$. Decreasing l_i till l_{min} results in the same per diem rate p^I for b_i/l_{min} patients, and the total revenue does not change. In other words, the hospital's monetary gain is zero and $G_i = 0$.

Case 2. $l_{0.25} < l_i \leq \bar{l}$

$R_i = p\big(1.15l_{0.25}+c(l_i-l_{0.25})\big)b_i/l_i$. The potential revenue under l_{min} is $1.15pb_i$. Therefore, the monetary gain from decreasing l_i to l_{min} is computed as: $G_i = 1.15pb_i - R_i = pb_i(1.15-c) - pl_{0.25}\big(1.15-c\big)b_i/l_i$.

Case 3. $\bar{l} < l_i \leq \bar{l}+2d$

$R_i = p\big(1.15l_{0.25}+c(\bar{l}-l_{0.25})+0.85c(l_i-\bar{l})\big)b_i/l_i$. $G_i = pb_i(1.15-c) - p\big(l_{0.25}(1.15-c)+0.15c\bar{l}\big)b_i/l_i$.

[34] Computation assumes that $p^I = 1.15\bar{p}$, $p^{III} = 0.85\bar{p}^{II}$, and areas A and B on Figure 1 are equal.

Case 4.

Hospital is financed according to a fee-for-service system, with per diem flat rate of p. The per patient revenue is pl_i, the number of patients is b_i/l_i and the total revenue is pb_i. If the length of stay becomes l_{min} under the per diem prospective payment system, hospital receives per diem rate $1.15p$ and total revenue $1.15pb_i$. So $G_i = 0.15pb_i$.

Since p is the same for each hospital, it may be ignored in defining the treatment variable in cases 1 through 4.

To test for heterogeneous effects, we interact D with the pre-reform values of quality measures or length-of-stay for the U.S. and Japanese reforms, respectively.

E.3 Empirical approach

We use the Semykina and Wooldridge (2010) and the Wooldridge (1995) approaches to account for sample selection and endogeneity in estimating panel data models. It should be noted that sample selection is observed both with the U.S. and Japanese reform participation. Indeed, our analysis with the U.S. data shows that although the value-based purchasing was a compulsory reform for all Medicare's hospitals, 14% of eligible hospitals did not join in 2013. Non-participation is associated with smaller potential annual monetary amount of punishment/reward,[35] which may be linked to hospital's size and proxied by the number of beds. Regarding the Japanese inpatient prospective payment reform, self-selection in terms of submitting the data to the MHLW may be associated with data-management practices at hospitals. Consequently, the binary variable for participation in the Japanese residency matching program may be an instrument for the fact of submitting the data, as coding of diagnoses is a prerequisite for qualifying to be a teaching hospital.

The model combines a selection equation for a latent variable h^*

$$h_{it}^* = z_{it}\gamma + \eta_i + \nu_{it}, h_{it} = I(h_{it}^* > 0) \tag{E.1}$$

and an intensity equation for the dependent variable y_{it}, observed under $h_{it} = 1$

$$y_{it} = G_{it}\delta_1 + G_{it}y_{i,t-1} + x_{it}\beta + \alpha_i + \zeta_{it} \tag{E.2}$$

An inclusion of the interaction term $G_{it}y_{i,t-1}$ allows estimating differential treatment effects, based on the pre-reform performance. The term leads to endogeneity in panel data estimates (given available data for the Japanese reform), owing to serial correlation between y_{it} and $y_{i,t-1}$.

[35] We inferred the potential amount from the size of hospital adjustment coefficient, which are prospectively announced based on historic data

References

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58:277–297.

Arellano, M. and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, 68:29–51.

Baron, D. P. and Myerson, R. B. (1982). Regulating a monopolist with unknown costs. *Econometrica: Journal of the Econometric Society*, 50(4):911–930.

Besstremyannaya, G. (2011). Managerial performance and cost efficiency of Japanese local public hospitals. *Health Economics*, 20(S1):19–34.

Besstremyannaya, G. (2015). Heterogeneous effect of residency matching and prospective payment on labor returns and hospital scale economies. Stanford Institute for Economic Policy Research, Discussion Paper 015-001, available at http://www-siepr.stanford.edu/RePEc/sip/15-001.pdf.

Besstremyannaya, G. and Shapiro, D. (2012). Heterogeneous effect of prospective payment system on hospital volume and quality. CEFIR Working Paper 181.

Bhattacharya, J., Vogt, W., Yoshikawa, A., and Nakahara, T. (1996). The utilization of outpatient medical services in Japan. *The Journal of Human Resources*, 31(2):450–476.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87:115–143.

Borah, B. J., Rock, M. G., Wood, D. L., Roellinger, D. L., Johnson, M. G., and Naessens, J. M. (2012). Association between value-based purchasing score and hospital characteristics. *BMC health services research*, 12(1):464.

Busse, R. and Schwartz, F. (1997). Financing reforms in the German hospital sector: from full cost cover principle to prospective cases. *Medical Care*, 35(10):OS40–OS49.

Campbell, J. and Ikegami, N. (1998). *The Art of Balance in Health Policy. Maintaining Japan's Low-Cost, Egalitarian System.* Cambridge University Press: Cambridge.

Campbell, S. M., Reeves, D., Kontopantelis, E., Sibbald, B., and Roland, M. (2009). Effects of pay for performance on the quality of primary care in England. *New England Journal of Medicine*, 361(4):368–378.

Canay, I. (2011). A simple approach to quantile regression for panel data. *The Econometrics Journal*, 14(3):368–386.

Chalkley, M. and Malcomson, J. M. (1998). Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics*, 17(1):1–19.

Chalkley, M. and Malcomson, J. M. (2000). Government purchasing of health services. *Handbook of Health Economics*, 1:847–890.

Chen, L. M., Jha, A. K., Guterman, S., Ridgway, A. B., Orav, E. J., and Epstein, A. M. (2010). Hospital cost of care, quality of care, and readmission rates: penny wise and pound foolish? *Archives of Internal Medicine*, 170(4):340–346.

Chernozhukov, V. and Hansen, C. (2004). Instrumental variable quantile regression. Working Paper, available at http://www.mit.edu/∼vchern.

Chernozhukov, V. and Hansen, C. (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525.

Chernozhukov, V. and Hansen, C. (2008). Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142(1):379–398.

Christianson, J. B. and Conrad, D. (2011). Provider payment and incentives. *The Oxford Handbook of Health Economics*, pages 624–648.

Damberg, C. L., Raube, K., Teleki, S. S., and dela Cruz, E. (2009). Taking stock of pay-for-performance: a candid assessment from the front lines. *Health Affairs*, 28(2):517–525.

Damberg, C. L., Raube, K., Williams, T., and Shortell, S. M. (2005). Paying for performance: implementing a statewide project in California. *Quality Management in Healthcare*, 14(2):66–79.

Doran, T., Fullwood, C., Kontopantelis, E., and Reeves, D. (2008). Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework. *The Lancet*, 372(9640):728–736.

Eijkenaar, F., Emmert, M., Scheppach, M., and 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. and McGuire, T. (1986). Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of Health Economics*, 5(2):129–151.

Ellis, R. P. and McGuire, T. G. (1996). Hospital response to prospective payment: moral hazard, selection, and practice-style effects. *Journal of Health Economics*, 15(3):257–277.

Farrar, D. E. and Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, pages 92–107.

Fetter, R. B. and Freeman, J. L. (1986). Diagnosis related groups: product line management within hospitals. *Academy of Management Review*, 11(1):41–54.

Fujii, M. and Reich, M. (1988). Rising medical costs and the reform of japan's health insurance system. *Health Policy*, 9:9–24.

Galvao, A. (2011). Quantile regression for dynamic panel data with fixed effects. *Journal of Econometrics*, 164(1):142–157.

Grabowski, D., Afendulis, C., and McGuire, T. (2011). Medicare prospective payment system and the volume and intensity of skilled nursing facility services. *Journal of Health Economics*, 30:675–684.

Hamada, H., Sekimoto, M., and Imanaka, Y. (2012). Effects of the per diem prospective payment system with DRG-like grouping system (DPC/PDPS) on resource usage and healthcare quality in Japan. *Health Policy*, 107:194–201.

Hamilton, J. (1994). *Time Series Analysis*, volume 2. Princeton University Press, Princeton.

Hayashida, K., Imanaka, Y., Otsubo, T., Kuwabara, K., Ishikawa, K., Fushimi, K., Hashimoto, H., Yasunaga, H., Horiguchi, H., Anan, M., Fujimori, K., Ikeda, S., and Matsuda, S. (2009). Development and analysis of a nationwide cost database of acute-care hospitals in Japan. *Journal of Evaluation in Clinical Practice*, 15:626–633.

HCAHPS online (2013). HCAHPS fact sheet (HCAHPS Hospital Survey) August 2013. http://www.hcahpsonline.org/files/August_2013_HCAHPS_Fact_Sheet3.pdf.

Hirose, M., Imanaka, Y., Ishizaki, T., and Evans, E. (2003). How can we improve the quality of health care in Japan? Learning from JCQHC hospital accreditation. *Health Policy*, 66:29–49.

Hodgkin, D. and McGuire, T. (1994). Payment level and hospital response to prospective payments. *Journal of Health Economics*, 13(1):1–29.

Holmstrom, B. and Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 7:24–52.

Houle, S. K., McAlister, F. A., Jackevicius, C. A., Chuck, A. W., and Tsuyuki, R. T. (2012). Does performance-based remuneration for individual health care practitioners affect patient care? A systematic review. *Annals of Internal Medicine*, 157(12):889–899.

Ikegami, N. (1991). Japanese healthcare: low cost through regulated fees. *Health Affairs*, 10:87–109.

Ikegami, N. (2005). Medical care in Japan. *New England Journal of Medicine*, 133(19):1295–1299.

Ikegami, N. (2006). Should providers be allowed to extra bill for uncovered services? Debate, resolution, and sequel in Japan. *Journal of Health Politics, Policy and Law*, 31(6):1129–1149.

Ikegami, N. (2009). Games policy makers and providers play: introducing case-mix-based payment to hospital chronic care units in Japan. *Journal of Health Politics, Policy and Law*, 34(3):361–380.

Ikegami, N. and Campbell, J. (1995). Medical care in Japan. *New England Journal of Medicine*, 133:1295–1299.

Ikegami, N., Yoo, B., Hashimoto, H., Matsumoto, M., Ogata, H., Babazono, A., Watanabe, R., Shibuya, K., Yang, B., Reich, M., and Kobayashi, Y. (2011). Japanese universal health coverage: evolution, achievements, and challenges. *Lancet*, 378:1106–1115.

Ishikawa, K., Yamamoto, M., Kishi, D., and Nabeshima, T. (2005). New prospective payment system in Japan. *American Journal of Health System Pharmacy*, 62:1617–1619.

Joskow, P. L. and Rose, N. L. (1989). The effects of economic regulation. *Handbook of Industrial Organization*, 2:1449–1506.

Joskow, P. L. and Schmalensee, R. (1986). Incentive regulation for electric utilities. *Yale Journal on Regulation*, 4:1–49.

Kahn, C. N., Ault, T., Isenstein, H., Potetz, L., and Van Gelder, S. (2006). Snapshot of hospital quality reporting and pay-for-performance under Medicare. *Health Affairs*, 25(1):148–162.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94(448):1296–1310.

Kondo, A. and Kawabuchi, K. (2012). Evaluation of the introduction of a diagnosis procedure combination system for patient outcome and hospitalisation charges for patients with hip fracture or lung cancer in Japan. *Health Policy*, 107:184–193.

Kondo, A. and Shigeoka, H. (2013). Effects of universal health insurance on healthcare utilization, and supply-side responses: evidence from Japan. *Journal of Public Economics*, 99:1–23.

Kridel, D. J., Sappington, D. E., and Weisman, D. L. (1996). The effects of incentive regulation in the telecommunications industry: A survey. *Journal of Regulatory Economics*, 9(3):269–306.

Kritzer, H. M. (1976). Problems in the use of two stage least squares: standardization of coefficients and multicollinearity. *Political Methodology*, pages 71–93.

Kuwabara, K., Imanaka, Y., Matsuda, S., Fushimi, K., Hashimoto, H., Ishikawa, K., and Horiguchi, H. (2006). Profiling of resource use variation among six diseases treated at 82 Japanese special functioning hospitals, based on administrative data. *Health Policy*, 78(2):306–318.

Kuwabara, K., Imanaka, Y., Matsuda, S., Fushimi, K., Hashimoto, H., Ishikawa, K., Horiguchi, H., Hayashida, K., and Fujimori, K. (2008). The association of the number of comorbidities and complications with length of stay, hospital mortality and LOS high outlier, based on administrative data. *Environmental Health and Preventive Medicine*, 13(3):130–137.

Kuwabara, K., Matsuda, S., Fushimi, K., Ishikawa, K., Horiguchi, H., Hayashida, K., and Fujimori, K. (2011). Contribution of case-mix classification to profiling hospital characteristics and productivity. *International Journal of Health Planning and Management*, 26:e138–e150.

Laffont, J. and Tirole, J. (1993). *A Theory of Incentives in Procurement and Regulation*. MIT Press.

Laffont, J.-J. and Tirole, J. (1986). Using cost observation to regulate firms. *Journal of Political Economy*, pages 614–641.

Laffont, J.-J. and Tirole, J. (1990). The regulation of multiproduct firms: Part I: Theory. *Journal of Public Economics*, 43(1):1–36.

Lindenauer, P. K., Remus, D., Roman, S., Rothberg, M. B., Benjamin, E. M., Ma, A., and Bratzler, D. W. (2007). Public reporting and pay for performance in hospital quality improvement. *New England Journal of Medicine*, 356(5):486–496.

Ma, C. A. (1994). Health care payment systems: cost and quality incentives. *Journal of Economics & Management Strategy*, 3(1):93–112.

Ma, C. A. (1998). Health-care payment systems: cost and quality incentives–Reply. *Journal of Economics & Management Strategy*, 7(1):139–142.

Mannion, R., Marini, G., and Street, A. (2008). Implementing payment by results in the English NHS: Changing incentives and the role of information. *Journal of Health Organization and Management*, 22(1):79–88.

Matsuda, S., Ishikawa, K., Kuwabara, K., Fujimori, K., Fushimi, K., and Hashimoto, H. (2008). Development and use of the Japanese case-mix system. *Eurohealth*, 14(3):25–30.

McKnight, R. (2006). Home care reimbursement, long-term care utilization, and health outcomes. *Journal of Public Economics*, 90:293–323.

Milgrom, P. and Roberts, J. (1995). Complementarities and fit strategy, structure, and organizational change in manufacturing. *Journal of Accounting and Economics*, 19(2):179–208.

Milgrom, P. and Shannon, C. (1994). Monotone comparative statics. *Econometrica*, pages 157–180.

Ministry of Health, Labor and Welfare (2012a). Heisei 24nendo sinryouhoushuu kaitei-niokeru DPC seido (DPC/PDPS)-no taiyou-nitsuite (gaiyou) (An outline of DPC/PDPS system within the 2012 revision of fee schedule, Mar 28).

Ministry of Health, Labor and Welfare (2012b). Heisei 24nendo sinryouhoushuu kaitei-niokeru DPC seido (DPC/PDPS)-no taiyou-nitsuite (hosoku jikou) (Supplementary information on DPC/PDPS system within the 2012 revision of fee schedule, Apr 25).

Ministry of Health, Labor and Welfare (2013). Heisei 26nendo sinryouhoushuu kaiteino sukejuuru (Schedule for the 2014 revision of the fee schedule).

Ministry of Health, Labor and Welfare (2014a). DPC/PDPS shoubyouna koodingu tekisuto (Handbook for coding illnesses in DPC/PDPS, Mar 26).

Ministry of Health, Labor and Welfare (2014b). Heisei 26nendo sinryouhoushuu kaitei-niokeru DPC seido (DPC/PDPS)-no taiyou-nitsuite (gaiyou) (An outline of DPC/PDPS system within the 2014 revision of fee schedule, Mar 26).

Miraldo, M., Siciliani, L., and Street, A. (2011). Price adjustment in the hospital sector. *Journal of Health Economics*, 30(1):112–125.

Moreno-Serra, R. and Wagstaff, A. (2010). System-wide impacts of hospital payment reforms: evidence from Central and Eastern Europe and Central Asia. *Journal of Health Economics*, 29(4):585–602.

National Institute of Population and Social Security Research (2005). Nihon shakai hoshou shiryou 1980-2000 (Materials on social security in Japan in 1980-2000).

Nawata, K. and Kawabuchi, K. (2013). Evaluation of the dpc-based inclusive payment system in japan for cataract operations by a new model. *Mathematics and Computers in Simulation*, 93:76–85.

Okamura, S., Kobayashi, R., and Sakamaki, T. (2005). Case-mix payment in Japanese medical care. *Health Policy*, 74:282–286.

Parente, P. M. and Santos Silva, J. (2015). Quantile regression with clustered data. *Journal of Econometric Methods, DOI: 10.1515/jem-2014-0011.*

Pearson, S. D., Schneider, E. C., Kleinman, K. P., Coltin, K. L., and Singer, J. A. (2008). The impact of pay-for-performance on health care quality in Massachusetts, 2001–2003. *Health Affairs*, 27(4):1167–1176.

Rosenthal, M. B. (2008). Beyond pay for performance–emerging models of provider-payment reform. *New England Journal of Medicine*, 359(12):1197–1200.

Rosenthal, M. B., Fernandopulle, R., Song, H. R., and Landon, B. (2004). Paying for quality: providers' incentives for quality improvement. *Health Affairs*, 23(2):127–141.

Rosenthal, M. B., Frank, R. G., Li, Z., and Epstein, A. M. (2005). Early experience with pay-for-performance: from concept to practice. *JAMA*, 294(14):1788–1793.

Ryan, A. M. and Blustein, J. (2011). The effect of the MassHealth hospital pay-for-performance program on quality. *Health Services Research*, 46(3):712–728.

Ryan, A. M., Blustein, J., and Casalino, L. P. (2012). Medicare's flagship test of pay-for-performance did not spur more rapid quality improvement among low-performing hospitals. *Health Affairs*, 31(4):797–805.

Semykina, A. and Wooldridge, J. M. (2010). Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics*, 157(2):375–380.

Shleifer, A. (1985). A theory of yardstick competition. *RAND Journal of Economics*, 16:319–327.

Sood, N., Buntin, M., and Escarce, J. (2008). Does how much and how you pay matter? Evidence from the inpatient rehabilitation care prospective payment system. *Journal of Health Economics*, 27:1046–1059.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4).

Thompson, J. D., Averill, R. F., and Fetter, R. B. (1979). Planning, budgeting, and controlling–one look at the future: case-mix cost accounting. *Health Services Research*, 14(2):111–125.

Wang, H. and He, X. (2007). Detecting differential expressions in GeneChip microarray studies: a quantile approach. *Journal of the American Statistical Association*, 102(477):104–112.

Werner, R. M. and Dudley, R. A. (2012). Medicare's new hospital value-based purchasing program is likely to have only a small impact on hospital payments. *Health Affairs*, 31(9):1932–1940.

Werner, R. M., Kolstad, J. T., Stuart, E. A., and Polsky, D. (2011). The effect of pay-for-performance in hospitals: lessons for quality improvement. *Health Affairs*, 30(4):690–698.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics*, 126:25–51.

Wooldridge, J. (2007). Quantile methods. NBER Summer Institute, Lecture Notes 14, available at http://www.nber.org/WNE/lect_14_quantile.pdf.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, 68(1):115–132.

Yasunaga, H., Ide, H., Imamura, T., and Ohe, K. (2005). Impacts of the Japanese diagnosis procedure combination based payment system on cardiovascular medicine-related costs. *International Heart Journal*, 46:855–866.