Financial incentives to encourage value-based health care

Anthony Scott, Miao Lui, Jongsay Yong

*Anthony Scott, PhD a.scott@unimelb.edu.au
Miao Liu miao.liu@unimelb.edu.au
Jongsay Yong, PhD jsyong@unimelb.edu.au

Melbourne Institute of Applied Economic and Social Research
The University of Melbourne
Parkville
Victoria 3010
Australia

*Corresponding author

Keywords: financial incentives, pay for performance, value-based health care

Acknowledgements. This review was funded by a grant from Medibank Private Health Research Fund. The views are those of the authors and not Medibank Private. The dataset is available from the authors on request.

Declaration of conflicts of interest. The Authors declare that there is no conflict of interest

Financial incentives to encourage value-based health care.

Abstract

This paper reviews the literature on the use of financial incentives to improve the provision of value-based health care. Eighty studies of 44 schemes from 10 countries were reviewed. The proportion of positive and statistically significant outcomes was close to 0.5. Stronger study designs were associated with a lower proportion of positive effects. There were no differences between studies conducted in the United States compared to other countries; between schemes that targeted hospitals or primary care; or between schemes combining pay for performance with rewards for reducing costs, relative to pay for performance schemes alone. Paying for performance improvement is less likely to be effective. Allowing payments to be used for specific purposes, such as quality improvement, had a higher likelihood of a positive effect, compared to using funding for physician income. Finally, the size of incentive payments relative to revenue was not associated with the proportion of positive outcomes.

INTRODUCTION

Encouraging health care providers to provide high value health services is now a central component of policy for US Medicare and is being adopted in many other countries. There have been a number of developments that have led to the focus on value-based health care. Continuing rises in health expenditures are not sustainable, especially in the face of existing policies which do not seem to have reduced expenditure growth or increased performance. In addition, there is growing use of pay for performance for health care providers which is usually accompanied by the measurement and public reporting of provider performance.

'Top down' approaches by governments and insurers are also being complemented by physicians who are beginning to grasp the importance of value based health care. The Choosing Wisely campaign and emphasis on over-diagnosis has led to a public recognition by the medical profession in many countries that some services they currently provide should not be provided (Malhotra et al. 2015; O'Callaghan, Meyer, and Elshaug 2015). Together, these developments could be essential building blocks in reducing waste and improving efficiency in health care while improving quality and outcomes within the resources available (Berwick and Hackbarth 2012).

A key driver that governments and insurers are using to encourage the provision of value-based health care is changes to funding and financial incentives. Value-based purchasing (VBP) or pricing or commissioning refers to a broad set of related policies that measure and pay for health care services defined in terms of their relative quality, outcomes and cost. This definition is used and applied quite broadly in the literature to include pay for performance schemes as well as schemes that focus on both quality and cost through shared savings models that are common in Accountable Care Organisations (ACOs) in the US. The move to

value-based health care in the US signals a clear move away from fee-for-service 'volume' based health care towards payment and rewards based on quality improvement and cost savings. The Affordable Care Act has led to a major shift to value-based purchasing in the US Medicare system. The intention is, *"to have 85% of all Medicare fee-for-service payments tied to quality or value by 2016"*, and *"90% by 2018"* and *"to have 30% of Medicare payments tied to quality or value through alternative payment models by the end of 2016 and 50% of payments by 2018"* (Burwell 2015).

ACOs were introduced by the affordable care act into Medicare, (Pioneer ACOs' and Medicare Shared Savings Programs) and a range of private ACOs were also established, some of which were built on existing Primary Care Medical Home models of care (PCMHs). The key innovation of some ACOs is the combination of pay for performance as well as shared saving models, where providers share in any savings (one-sided models) and are at risk for deficits (two-sided models). The shared savings aspect of ACOs typically includes a spending target for the following year (based on historical spending trends and morbidity). If spending per patient ends up below this target the organisation/group receives a share of these savings though this depends on meeting specific quality thresholds. If spending exceeds the target, the organisation may need to bear some of this deficit. ACOs in the US cover over 18 million people, including over 360 ACOs funded by Medicare (Song 2014). Their key features are that they are groups of physicians and hospitals responsible for both the cost and quality of care for their patients, and with a strong primary care focus (Colla and Fisher 2014).

There have been a number of systematic reviews of the use of pay for performance for health care providers (Scott et al. 2011; Witter, Fretheim, Kessy, and Lindahl 2012; Langdown and

Peckham 2014; Van Herck et al. 2010; Damberg et al. 2014; Houle, McAlister, Jackevicius, Chuck, and Tsuyuki 2012), and reviews of those reviews (Eijkenaar, Emmert, Scheppach, and Schoffski 2013; Flodgren et al. 2011). Most come to similar conclusions that the effect of pay for performance is mixed, that there is a lack of evaluations of the cost-effectiveness of these schemes, and that there are few well-designed empirical studies.

Previous literature has also generated discussion around the appropriate design principles of such schemes (Conrad and Perry 2009; Eijkenaar 2013; Scott et al. 2010; Scott and Harris 2012; McKethan and Jha 2014; Glasziou et al. 2012; Rosenthal and Frank 2006). Poor design is partly due to the lack of a theoretical/conceptual framework as to why the particular design used is likely to influence behaviour. The size of the incentives as a proportion of revenue/income is also a key feature of earlier designs for which these schemes were found to have very low power to motivate behaviour change. A key issue is also the salience of incentives or the extent to which physicians or clinical teams were aware of the rewards, or were able to use rewards either as direct supplements to their income or for quality improvement activities. A further issue is that with tournament designs, where providers are rewarded only if they meet a target threshold of the indicator or are ranked in the top of the distribution of performance, providers who already meet the threshold will be rewarded but have not actually changed their behaviour. Low performing providers are unlikely to change their behaviour because the threshold is too high. In contrast, paying for improvements in performance between two time points and relative to own baseline performance would then provide incentives for low as well as high performers to change their behavior. These issues have been cited as reasons why some schemes do not seem to have improved performance, but there remains little systematic evidence on appropriate design features.

The aim of this paper is to review the recent (since 2010) empirical literature that evaluates the effect of value-based purchasing in health care. The objectives of the review are twofold.

i) To update the findings from previous reviews in the context of value-based purchasing by summarising results from existing evaluations in terms of the reported impact of schemes on cost and quality.
ii) To summarise the key design features of these schemes.

METHODS

Studies were included if they evaluated interventions that examined only pay for performance, or if the interventions included incentives for performance combined with incentives for reducing costs. The Alternative Payment Model (APM) framework is now being used to classify different types of payment schemes in the United States, with four levels of classification (APM FPT Work Group 2016). Category 1 is fee for service payment with no ties to quality. Category 2 is fee-for-service payment with pay for performance. Category 3 includes Category 2 plus a budget or cost target with either one-sided shared savings (Category 3A) or two-sided (Category 3B) risk sharing. This category is limited to where services/procedures are incentivised (e.g. hip fracture, immunisation), rather than conditions or populations. Category 4 is Category 3 but where fee-for-service doesn't exist and has been replaced by some type of capitation payment to a defined population (4A is a population defined by a medical condition, eg asthma or diabetes; 4B is for a population defined by geography or enrolment and includes vertically integrated systems). Salaried payment is mentioned in 4B.

In this review we focus on studies that introduce pay for performance (Category 2), and studies that include pay for performance as well as rewards for cost reductions (Categories 3 and 4). Movements from Category 1 to Category 4 imply greater accountability for costs and quality and a greater focus on population health management (Rajkumar, Conway, and Tavenner 2014). Studies that included incentives only for reducing costs, such as capitation payment, were not included in this review.

We searched PubMed and EconLit electronic databases for journal articles with the keywords of "value-based purchasing", "pay for performance", and "accountable care organisations". Articles were also found through screening the citations of included papers and previous reviews, forward citation tracking of previous reviews and included studies, searching websites of authors frequently publishing in the field, and grey literature searches of key websites. The search was conducted between March and July 2015. We included studies in English from high income and low and middle income countries. We included studies published between 2010 until July 2015.

Studies using qualitative data, reviews of the literature, editorials, and opinion pieces were excluded from the review. Studies were included if they examined the impact of schemes on any type of outcome (e.g. costs, utilisation, expenditures, quality of care, health outcomes). Studies were included if incentives were targeted at individual or groups of medical practitioners or hospitals. In terms of study designs, observational and modelling studies with no control group were excluded, as were before and after studies and controlled before and after studies that only compared the means of primary outcomes with no attempt to adjust for confounders. Before and after studies and controlled before and after studies that used regression analysis to attempt to control for confounders were included. Studies that used

interrupted time series designs (ITS), difference-in-difference (DID) designs, and randomised controlled trials (RCTs) were included.

Initial review of titles and abstracts were used to select papers, supplemented with full text review if the abstract contained insufficient information or where there was uncertainty. The above criteria were applied by one author (ML) with borderline papers discussed with another (AS). All included papers were reviewed again for inclusion by AS.

Data extracted on whether the incentive scheme had an impact.

It was not possible to summarise and synthesise actual effect sizes due to substantial heterogeneity in the types of outcomes between studies and because of variation in the number and different types of outcomes included within each study. A simple vote count was used to summarise the effects of the interventions. Vote counting is useful to summarise whether there is likely to be an effect or not and gives an estimate of the probability that the intervention will be effective. This is recommended by the Cochrane Collaboration where there is no consistent outcome measure across studies. Furthermore, it is recommended that vote counting should ignore statistical significance when classifying positive or negative effects as these are often influenced by unit of analysis errors or small sample sizes (Higgins and Green 2011). However, this could lead to upward bias in the reporting of studies with positive effects, and so we also note any unit of analysis errors and sample size issues. We report three measures summarising the likelihood of an overall effect. First, the total number of outcome measures across and within all included studies is counted, and the proportion which had a statistically significant and positive effect (noting issues of unit of analysis error and small sample size). Statistical significance is defined at the 10% level. Outcomes aggregated to the level of the published paper (study) and to the level of the incentive scheme

are also reported. We calculate the mean proportion of positive and statistically significant outcomes per scheme and per study (Bushman and Wang 1994).

Data extracted on characteristics of studies and schemes.

Included papers were summarised with respect to their country, setting (primary care and/or hospitals), author-reported objectives, data (sample size, time period, unit of analysis), study design, and author-reported results. We extract the following data on specific design features of the schemes:

i) country where the scheme was conducted,
ii) whether the scheme was targeted at hospitals or primary care,
iii) whether the scheme included rewards for quality (APM Category 2) or for both quality and costs (APM Categories 3 and 4),
iv) whether incentive payments could be used as physician income, whether there was discretion as to their use, or whether they could only be used for specific purposes (but not physician income),
v) whether the scheme provided incentives for performance improvement rather than meeting a single threshold (including studies which had more than one threshold, or directly paid for a change in performance between two time points).
vi) the percentage size of the incentive payment relative to total revenue or income, and;
vii) study design: before and after studies and controlled before and after studies that used regression analysis to attempt to control for confounders; ITS; DID; and RCTs.

Statistical analysis.

Associations between the above characteristics of the incentive scheme (or study) with the proportion of positive outcomes are examined. Regression analysis is conducted using the

study as the unit of analysis, and with the proportion of positive outcomes per study as the dependent variable. Independent variables include study and scheme characteristics, described and coded as above. As the dependent variable is a proportion bounded between zero and one, and with observations at the bounds, a fractional logit is estimated using a generalised linear model with a logit link function (Papke and Wooldridge 1993). The data have a hierarchical structure as for many schemes, several studies have been conducted. To account for the correlation of residuals of studies within schemes due to unobserved scheme-level characteristics, the observations are clustered by scheme. Unlike a random effects model which assumes that the within-cluster correlation is correctly specified and is the same for each scheme, the clustering of standard errors allows for error correlations within each scheme to vary arbitrarily (Cameron and Miller 2015).

RESULTS

After 166 duplicates were removed, 1,269 papers underwent an initial title and abstract review. This led to a further 1,049 papers being excluded because they were literature reviews, descriptive studies, qualitative studies, editorials or studies that were not in English. The full text of the remaining 220 papers was reviewed according to the selection criteria. Of these, a further 140 papers were excluded, leaving 80 empirical studies included in the review (Figure 1).

In which countries and settings were the schemes and studies?

The 80 included studies were conducted across 10 countries. Forty four separate schemes were evaluated (Table 1): 25 (57%) schemes were in the US, five (11%) were in the UK National Health Service (NHS), and three (7%) from Taiwan. China, Canada, Italy and Australia each had two schemes evaluated, and France, the Philippines, and Rwanda had one

scheme each. Forty two studies (53%) were conducted on US schemes, followed by 16 (20%) in the UK, and nine (11%) in Taiwan (Table 1).

Hospitals were the setting for 13/44 (29.5%) schemes. The majority of schemes included incentives targeted at multi-specialty physician groups or primary care physicians. For some schemes where the setting of the physicians was not clear, the setting was defined using the characteristics of patients, for example patients receiving dialysis were classified as hospitals, and immunisation was classified as primary care.

Of the 25 schemes in the US (42 studies), fifteen schemes were implemented by a variety of local private insurers or networks. Nine schemes were run by Medicare or Medicaid. Of the 15 private schemes, the Alternative Quality Contract (AQC) implemented by Blue Cross Blue Shield in Massachusetts was examined in eight studies. These studies examined the effects on primary care physicians' costs and quality after one year (Song et al. 2011), two years (Song et al. 2012), and four years (Song et al. 2014). Other studies focused on different aspects of the scheme, such as spending and utilisation (McWilliams, Landon, and Chernew 2013), emergency department use (Sharp, Song, Safran, Chernew, and Mark Fendrick 2013), spending and utilisation of specific technologies (Song, Fendrick, Safran, Landon, and Chernew 2013), paediatric spending and quality (Chien et al. 2014), and pharmaceutical spending (Afendulis et al. 2014).

Of the fifteen different schemes run by Medicare or Medicaid, five studies evaluated the Premier Hospital Quality Incentive Demonstration Program which began in 2003. Jha, Orav, and Epstein (2010) examined the impact on poor patients, Kruse, Polsky, Stuart, and Werner (2012) examined the effects on costs and revenues, and Ryan, Blustein, and Casalino (2012)

examined the impact of a change in the design of the scheme in 2006 to encourage higher performance from low performing hospitals. One study examined the longer term impact of the scheme (Jha, Joynt, Orav, and Epstein 2012), and another examined the longer term impacts on surgical outcomes after the change in design in 2006 (Shih, Nicholas, Thumma, Birkmeyer, and Dimick 2014). Two studies evaluated the effect of Medicare not funding 'never events' and hospital acquired conditions. Lee et al. (2012) examined the impact on two types of infections covered by the scheme and Kawai et al. (2015) examined effects on coding and billing practices.

Three studies evaluated the more recent Medicare ACOs (including Pioneer ACOs and the Medicare Shared Savings Program (MSSP)) targeted at physician medical groups, that also include shared savings programs in addition to pay for performance. One study examined changes in patient's experience (McWilliams, Landon, Chernew, and Zaslavsky 2014), another examined effects on cost and quality after one year (McWilliams, Chernew, Landon, and Schwartz 2015), and another after two years (Nyweide et al. 2015).

Three studies examined the impact of the Medicare Physician Group Practice Demonstration, and early version of an ACO that included 10 medical groups, which ran from 2005 to 2010. Colla et al. (2012) examined the impact of health expenditures during the first four years. A subsequent study focussed on health expenditures for cancer patients (Colla, Lewis, Gottlieb, and Fisher 2013), whilst the third study examined the impact on the utilisation of cardiovascular care (Colla et al. 2014). One study examined the new Medicare Hospital Value-Based Purchasing scheme during the first year of its implementation (Ryan, Burgess, Pesko, Borden, and Dimick 2015). This scheme allocates a proportion of hospital DRG payments according to quality and patient experience.

In the UK, five different schemes (reported in 16 studies) were examined: the Advancing Quality Initiative, Best Practice Tariffs, the Quality and Outcomes Framework, NHS Stop smoking services, and Clinically Directed Enhanced Services. Three studies examined the Advancing Quality Initiative for NHS hospitals focusing on the short term effects (Sutton et al. 2012), long term effects (Kristensen et al. 2014) and cost-effectiveness (Meacock, Kristensen, and Sutton 2014). Two papers examined the impact of Best Practice Tariffs for NHS hospitals: one was an empirical paper focussing on day surgery (Allen, Fichera, and Sutton 2016) and the other the final report which focused on a broader set of tariffs in addition to day surgery (McDonald et al. 2012).

Nine studies focussed on different aspects of the Quality and Outcomes Framework pay for performance scheme for GPs. Of these, two studies examined diabetes (Alshamsan, Lee, Majeed, Netuveli, and Millett 2012; Gallagher, Cardwell, Hughes, and O'Reilly 2015), two examined the management of hypertension (Simpson, Hannaford, Ritchie, Sheikh, and Williams 2011; Serumaga et al. 2011), one the prescribing of contraceptives (Arrowsmith, Majeed, Lee, and Saxena 2014), and one study examined disparities in hypertension, stroke and heart disease management (Lee, Netuveli, Majeed, and Millett 2011). Another study examined the removal of incentives for stroke, asthma, heart disease, diabetes and psychosis (Kontopantelis et al. 2014), one paper examined the impact on emergency admissions for ambulatory sensitive conditions (Harrison et al. 2014), and one examined the impact of an increase in the target payment threshold for influenza immunisation (Kontopantelis et al. 2012). Finally, one study examined a scheme for pay for outcomes for NHS stop smoking services (McLeod, Blissett, Wyatt, and Mohammed 2015), and another examined a local pay for performance scheme for alcohol screening and brief intervention (Hamilton et al. 2014).

In Taiwan, all nine studies focused on three schemes introduced by the National Health Insurance P4P program which introduced three disease-specific pay for performance schemes. Each scheme was different. Six studies were aimed at primary care physicians and focused on different outcomes in diabetes care (DM-P4P). Two of these (Hsieh, Tsai, Shin, Mau, and Chiu 2015; Tan, Pwu, Chen, and Yang 2014) were economic evaluations of the diabetes care scheme from the same authors but at early and later stages of the scheme. One study focused on emergency department visits (Yu, Tsai, and Kung 2014), one of self-care (Chen, Lee, and Kuo 2012), one examined effects after four years of the scheme (Cheng, Lee, and Chen 2012), another used data from one year of the scheme (Lai and Hou 2013). Three studies focussed on the scheme implemented in hospitals. One study focused on breast cancer screening (BC-P4P) (Kuo, Chung, and Lai 2011), and two on tuberculosis (TB-P4P) (Tsai et al. 2010; Li et al. 2010).

In Canada, three studies evaluated two schemes: an Emergency Department program in British Columbia (ED P4P) (Cheng and Sutherland 2013), and two studies focussed on different periods of a P4P program for family physicians in Ontario (Kiran, Wilton, Moineddin, Paszat, and Glazier 2014; Li, Hurley, DeCicca, and Buckley 2014). Of the three studies from China, two reported results on different aspects of the same scheme in rural clinics; outpatient and inpatient utilisation (Powell-Jackson, Yip, and Han 2014) and expenditures, patients satisfaction and antibiotic prescriptions (Yip et al. 2014).

Did the schemes work?

The 80 studies across 44 schemes reported on a total of 1,302 outcome measures, of which 46% (595) were positive and statistically significant at the 10% level (Table 2). There was a

wide range of outcome measures including health expenditures and quality of care measures and indicators, and outcomes from sub-group analysis. Each scheme had an average of 29.6 outcome measures, with 13.5 positive and statistically significant outcomes. The mean of the percentage of positive and statistically significant outcomes per scheme was 56%. Each of the 80 empirical studies reported an average of 16.3 outcome measures, of which 7.4 were positive and statistically significant. The mean percentage of positive outcomes per study was 54%.

Table 3 shows the ranking of schemes by the mean proportion of positive outcome measures. Only three schemes reported no impact on any of the outcomes, all from the US (Chien, Eastman, Li, and Rosenthal 2012; Ryan and Blustein 2011; Ryan, Burgess, Pesko, Borden, and Dimick 2015).

These data were summarised at country-level by taking the proportion for each scheme in Table 3, and calculating the mean for each country (Table 4). Of the 25 schemes in the US, an average of 56% of outcomes was positive and statistically significant. This was no different compared to the mean for all non-US schemes, which was also 56%. This compares to 91% for the P4P schemes in Taiwan, 75% in Canada and Italy, and 48% in the UK.

In the US, the Premier Hospital Quality Incentive Demonstration was one of the least effective schemes (Table 3). This adds to previous studies prior to 2010 which generally showed no effects of this scheme (Ryan 2009). Six out of eight studies of the Alternative Quality Contract showed an impact of the scheme on both reducing spending and improvements in quality after four years of the scheme. The evaluation focussed on the patients of primary care physicians, compared with a matched control group. After one year,

the difference-in-difference analysis found that the ACQ had reduced the growth in expenditures and improved quality of care for chronic disease patients and paediatrics (Song et al. 2011). At the end of the second year, reduced cost growth and quality improvements continued (Song et al. 2012). At this time there was evidence that there was a reduction in expenditure growth for Medicare patients who were not covered, but who were enrolled with the same provider organisations participating in the ACQ (McWilliams et al., 2013). This suggested some positive spillovers within these provider organisations, though there were no consistent improvements in quality of care amongst Medicare patients. A follow up analysis after four years to 2012 found evidence of continued savings and quality improvements (Song et al. 2014). The other two studies found no effect on emergency department use or pharmaceutical spending (Afendulis et al. 2014; Sharp, Song, Safran, Chernew, and Mark Fendrick 2013).

There are also positive effects from the three studies examining Medicare ACOs which showed some evidence of reductions in the growth of spending (McWilliams, Chernew, Landon, and Schwartz 2015; Nyweide et al. 2015), with evidence that patient experience was no worse than before (McWilliams, Landon, Chernew, and Zaslavsky 2014; Nyweide et al. 2015). None of these three studies examined effects on other measures of quality of care. Of the three studies that examined the Physician Practice Group Demonstration, two found evidence of reductions in spending (Colla, Lewis, Gottlieb, and Fisher 2013; Colla et al. 2012) and one found no evidence of a change in utilisation of cardiovascular care (Colla et al. 2014). One study on the recent Medicare Value-Based Purchasing Program for hospitals showed no effects on quality after one year (Ryan, Burgess, Pesko, Borden, and Dimick 2015).

Evidence suggests that not reimbursing hospitals for 'never' events', such as wrong site surgery and hospital acquired infections, is unlikely to reduce the incidence of such events (Lee et al. 2012a) and may only change coding practices such that these events are less likely to be coded and reported (Kawai et al. 2015). Such a scheme was modelled in public hospitals in Victoria and found that DRG funding between hospitals would be re-distributed such that hospital revenues would change between -2.5% to 1.8% (McNair P et al. 2009).

In the UK, of the three studies examining the Advancing Quality Initiative for NHS hospitals, one study found evidence of short term effects during the program's first 18 months (Sutton et al. 2012), but analysis after a further two years found these effects had disappeared (Kristensen et al. 2014). Meacock, Kristensen, and Sutton (2014) conducted a cost-utility analysis of the scheme based on the results from the first 18 months (Sutton et al. 2012) and found it to be cost-effective.

There was evidence of a positive effect on day surgery rates of Best Practice Tariffs for day surgery for cholecystectomy within the first year (Allen, Fichera, and Sutton 2016). There was no impact on quality of care for stroke (McDonald et al. 2012). For the hip fracture tariff there was a positive effect on the timing of hip fracture surgery, a fall in mortality rates, and an increase in the proportion of patients being discharged to their usual place of residence.

Of the three schemes (nine studies) conducted in Taiwan of the National Health Insurance P4P scheme, eight showed a positive effect. Six studies evaluated the impact of the scheme on diabetes care. One compared patients in and out of the scheme using a telephone survey, and found that those in the scheme had better levels of self-care after controlling for observed characteristics (Chen, Lee, and Kuo 2012). Cheng, Lee, and Chen (2012) found that diabetes

patients in the scheme had a higher utilisation of tests and visits initially, but that this difference fell over time. They also had lower diabetes-related hospitalisations. Overall, health expenditures were higher in the first year, but lower in subsequent years. Hsieh, Tsai, Shin, Mau, and Chiu (2015) and Tan, Pwu, Chen, and Yang (2014) conducted cost-effectiveness analyses early and later on in the scheme. Both studies found the scheme to be cost-effective. A further study found that patients in the program were more likely to receive guideline-recommended tests and examinations, and that this was also the case for patients not enrolled in the program but seeing the same physicians (Lai and Hou 2013). The study that showed a negative effect showed an increase in emergency admissions for diabetes patients (Yu, Tsai, and Kung 2014). Two studies examined tuberculosis treatment and found that the cure rate, length of treatment and default rates improved (Li et al. 2010; Tsai et al. 2010). Finally, one study examined breast cancer screening and found patients had improved quality of care, higher five-year survival rates, and lower rates of re-occurrence (Kuo, Chung, and Lai 2011).

The association between scheme/study characteristics and the effectiveness of each scheme.

To relate the proportion of positive outcomes to design features and study characteristics, we estimate two generalised linear models. The results are shown in Table 5. Model 1 shows results for a sample with 76 studies, as the variable on how the incentives were used contained some missing values. Model 2 shows the results for the full sample of 80 studies without this variable.

Study design. The study design had the most consistent and one of the largest effects on the proportion of positive outcomes across all models. Relative to the weakest designs (before and after studies and controlled before and after studies with regression), the percentage of

positive and statistically significant outcomes for other study designs was between 13 and 25 percentage points lower. Though we did not formally assess study quality, a number of general issues emerge. The risk of bias is generally much higher in the case-control, before and after, and controlled before and after studies. For example, all nine studies from Taiwan (mean percentage of positive outcomes of 91%) used a case-control or before and after design. The designs (especially case-control studies) cannot control for baseline differences or longer term trends in quality. Other studies of the Taiwan schemes have shown that there may have been substantial selection bias of patients enrolled in the program, such that any positive effects were likely to have been due to selection rather than the impact of the program (Chang, Lin, and Aron 2012; Chen, Chung, Lin, and Lai 2011). Some studies attempt to account for this by using regression analysis to control for differences in health status and other characteristics between the intervention and control groups whilst others use matching to help ensure that patients are similar in both groups. Several studies use fixed effects panel data analysis to control for unobserved patient-level factors. In these ways, they may partly control for factors influencing selection of patients into the scheme.

Generally, the ability of most studies to control for the selection of providers into incentive schemes is weak. There are likely to be unobserved factors which influence selection into schemes, including providers selecting into schemes, voluntary or otherwise, who may already be good performers. These issues have not been extensively studied.

There are a large number of difference-in-difference study designs – these were used in 43% of all studies. These designs seem to have proliferated in this literature, yet are largely absent from the Cochrane nomenclature. They are recognised as strong designs since they could control for observable factors, unobservable fixed effects, and general time trends in the

outcomes. The interpretation of these study results as causal effects depends on whether studies tested for whether the trends in primary outcomes were the same in the control and intervention groups before the intervention. Parallel pre-intervention trends help ensure that any differences in trends detected in the analysis are not a result of differences that existed before the intervention (Ryan, Burgess, and Dimick 2015). A further issue is the selection of the control group. For example, some studies use non-incentivised health conditions as controls, or other geographic areas. These groups may be different to the intervention group in observed or unobserved ways that could influence differences in trends.

Setting. In terms of the setting of schemes, schemes conducted in the US had around an eight percentage point lower percentage of positive outcomes compared to other countries, but this was not statistically significant. For schemes that targeted hospitals, the percentage of positive outcomes were between six and seven percentage points lower compared to schemes targeting primary care, though this was not statistically significant.

Rewards for quality and costs. In terms of incentive design, schemes that provided incentives for both costs and quality (APM Categories 3 and 4) had a mean percentage of positive outcomes around 10 percentage points higher than schemes providing incentives only for quality (ie P4P, APM Category 2), but this was not statistically significant. Overall, thirty six schemes (52 studies) were classified as Category 2 schemes: schemes that introduced P4P. Eight schemes (28 studies) were classified as either Category 3 or 4 (APM FPT Work Group 2016).

US schemes under the ACO banner have introduced incentives to reward increases in quality combined with rewards for slowing expenditure growth (shared savings) (APM Categories 3

and 4). The Quality and Outcomes Framework in the UK pays GPs for quality but a major proportion of GP remuneration in the UK is by capitation, where all savings are retained by GP practices and they are at risk for deficits – this is classified as APM Category 4B though savings retained are not explicitly linked to the achievement of minimum quality. This highlights the key innovation in US schemes: that rewards for costs and quality are linked to each other: such that the amount of shared savings returned to providers depends on quality improvements.

In the US, the Category 3 and 4 schemes included ACO models such as the Alternative Quality Contract in Massachusetts (Afendulis et al. 2014; Chien et al. 2014; McWilliams, Landon, and Chernew 2013; Song, Fendrick, Safran, Landon, and Chernew 2013; Song et al. 2014; Song et al. 2011; Song et al. 2012). The contract includes pay for performance linked to shared savings through a risk-adjusted global budget and is a two-sided model where providers are also at risk for deficits (Chernew, Mechanic, Landon, and Safran 2011). Others include Medicare ACOs, which include the one-sided Medicare Shared Savings Plans and the Medicare Pioneer ACOs (McWilliams et al. 2015, McWilliams et al. 2014, Nyweide et al. 2015) where providers have two-sided contracts. The Medicare Physician Group Practice Demonstration was an earlier example of a one-sided shared savings model (Colla et al. 2014; Colla, Lewis, Gottlieb, and Fisher 2013; Colla et al. 2012).

There were only two schemes with two-sided contracts (ACQ, and Medicare ACOs) and one scheme with a one-sided contract (Physician Group Practice Demonstration). In Table 3, the mean proportion of positive outcomes in the older one-sided model was 0.62, compared to 0.5 for the Medicare Pioneer ACOs, and 0.47 for the ACQ.

Other schemes from outside of the US were also classified as rewarding both quality and costs. Global budgets and population-based capitation funding are more common in other countries, but the incentives from risk sharing are usually much weaker, especially being at risk for deficits. This included the Quality and Outcomes Framework in the UK where GPs are paid by population-based capitation payment in addition to P4P through the QOF. The Emilia-Romagna P4P in Italy was introduced alongside an existing system of patient enrolment and capitation payment for GPs (Fiorentini, Iezzi, Lippi Bruni, and Ugolini 2011). In Canada, the ED P4P scheme was introduced on top of an already existing global budget for hospital emergency departments (ED) and salaried payment for nurses and physicians, though it is unclear the extent to which the EDs could share savings or were at risk of deficits (Cheng and Sutherland 2013). The Shandong scheme in China introduced a capitated global budget and pay for performance to replace existing fee-for-service payment (Sun et al. 2014). The Ningxia scheme also implemented similar changes (Powell-Jackson, Yip, and Han 2014; Yip et al. 2014).

Use of payments. No study reported data on how incentive payments were actually used, rather they reported the rules for the use of payments. Three schemes (three studies) did not adequately describe who received the payments and did not provide any information on how they were used (Esse, Serna, Chitnis, Johnson, and Fernandez 2013; Hung and Green 2012; McLeod, Blissett, Wyatt, and Mohammed 2015; Tsai et al. 2010). The remaining 77 studies, covering 41 schemes, described how the incentive payments could potentially have been used by describing rules for the use of the payments. If studies stated that a medical group, hospital or other organisation received the payments as general revenue, it was assumed that the organisation had discretion as to how to use the money. Forty one schemes reported these rules. Of these, 14 (34.2%) reported that the incentives were paid directly to individual

physicians as bonuses on top of salary, or supplements to fee-for-service payments. Twenty four schemes (58.5%) paid the incentive funding to organisations and did not state any particular rules for their use, having discretion as to how to use the funds.

Three schemes (7.3%) (five studies) specified a particular use that was not physician income. However, the way payments were used was only weakly associated with the mean proportion of positive outcomes (Table 5). Incentives that had a specific use had a mean proportion of positive outcomes which were 24 percentage points higher, compared to incentives used for physician income. This is quite a large effect, statistically significant at the 10% level. However, the size of this effect may be due to the small cell size for this category (five studies examining three schemes), which may artificially lead to large coefficients in logit models. These three schemes included the ED P4P scheme in Ontario (mean proportion of positive outcomes=1 from one study) which was targeted at emergency department funding where physicians were paid separately from the department budget (Cheng and Sutherland 2013). In another scheme used by Kaiser Permanente (mean proportion of positive outcomes=1 from one study), incentives were paid to the medical group but could not be used for physician income (Lester et al. 2010). The Advancing Quality Initiative for NHS hospitals (mean proportion of positive outcomes=0.35 from three studies) specified explicitly that the funds were meant for quality improvements by the relevant clinical teams (Sutton et al. 2012). However, in the later part of the scheme there was some evidence that hospitals had employed additional staff in the relevant hospital departments, but there seemed to be no automatic transfer of funds to clinical teams, as a case has to be made by teams for the funds to be transferred (Kristensen et al. 2014).

Rewards for performance improvement. A key design feature that has been identified is whether the scheme rewarded for an improvement in performance between two time points, rather than paying for attainment of single thresholds as in tournament-based pay. This includes schemes that directly measure a change in performance, as well as schemes that have more than one threshold so that providers can move to a higher threshold over time.

Thirteen of the 44 schemes (37/80 studies) reported using a design that provided incentives for performance improvement. The results from the regression show that the percentage of positive outcomes from these schemes was just over 20 percentage points lower compared to schemes that did not pay for performance improvement, and this is statistically significant. Several schemes measured and paid for direct changes in performance measured between two time points. These schemes included Medicare's Premier Hospital Quality Incentive Demonstration which changed its design in 2006 to pay for improvement as well as 'attainment' (Jha, Joynt, Orav, and Epstein 2012; Ryan, Blustein, and Casalino 2012; Shih, Nicholas, Thumma, Birkmeyer, and Dimick 2014). Ryan, Blustein, and Casalino (2012) and Shih, Nicholas, Thumma, Birkmeyer, and Dimick (2014) examined the impact of this change and found there was no impact on quality. However, this scheme had already been found to be largely ineffective (Ryan 2009). An evaluation of the more recent Hospital Value-Based Purchasing Program (Ryan, Burgess, Pesko, Borden, and Dimick 2015) uses incentives for both attainment and improvement for quality that are linked to DRG payments under the prospective payment system for hospitals. Payments are withheld and are re-distributed to hospitals based on their relative quality of care. The Advancing Quality Scheme in the UK NHS (Sutton et al. 2012; Kristensen et al. 2014) which was modelled on the US Premier scheme, also used a similar scheme of attainment and improvement payments. The first year used a basic tournament which was subsequently changed to attainment and improvement

payments. After the first 18 months the scheme was changed from a bonus system to a penalty system where a fixed proportion of each hospitals income was withheld and paid only if quality thresholds were met. These thresholds were based on performance in the first year of the scheme. The short term impact of the scheme was not sustained in the longer term. The Massachusetts Medicaid Program pay for performance scheme for hospitals also used attainment and improvement payments but with no impact on quality (Ryan and Blustein 2011). Finally, the Diabetes Care Project in Australia based part of its payments in practice-level changes in HbA1c relative to baseline measures taken before the scheme began (Department of Health 2015).

Size of payments. Of the 44 schemes included in the review, 22 (49%) reported the size of the payment relative to total revenue or relative to previous payments. These included the relative size of bonuses, the relative size of withholds (penalties), and the additional loading on specific DRG payments or fees. The size of the incentives in each scheme are summarised in Table 6. Seven schemes (16%) reported the percentage size of payments to be below 5%. Four schemes were in primary care, whilst the rest were for larger schemes in hospitals, including the Medicare Premier scheme, the recent Medicare Hospital Value Based Purchasing scheme, and the Advancing Quality Scheme in the UK. Only one hospital scheme reported a loading of greater than 5%: the use of Best Practice Tariffs in the UK NHS with a loading of 24% of the DRG payment (Allen, Fichera, and Sutton 2016). The remaining 15 schemes reported payments greater than 5% and all were for primary care or medical groups. Eight schemes reported payments between 5% and 10% of revenue, and seven schemes reported payments of greater than 10% and up to 30% of revenue. This latter group included two schemes from China that used withholding payments, where up to 30% of the budget was held back and paid back based on quality (Powell-Jackson, Yip, and Han 2014; Sun et al.

2014; Yip et al. 2014). In these cases the proportion actually withheld is likely to be less than the maximum reported in the studies.

Figure 2 shows that there appears to be no relationship between incentive size and the proportion of positive outcomes per scheme. This is confirmed by adding this variable to the regression model. This reduces the sample size to 32 studies (22 schemes) with the results showing a one percent increase in the size of the incentive leads to a 1.2 percentage point increase in the proportion of positive outcomes, but this is not statistically significant (results available from authors on request).

DISCUSSION

This paper reviewed the recent literature on the impact and characteristics of schemes that use financial incentives to improve the provision of value-based health care. The review included 80 empirical studies, with just over half from the US and the remainder from nine other countries. These studies examined 44 separate incentive schemes, including 26 from the US.

The probability of a positive and statistically significant effect was around a half. Previous reviews have not summarised the results numerically and have rather used a narrative synthesis approach which may be subject to bias, rather than vote counting or other quantitative synthesis. Though we have acknowledged the drawbacks with vote counting, it does give a more objective measure of overall impact. A Cochrane 'review of reviews' included four high quality systematic reviews (from 84 reviews identified, most of which did not report any numerical data) and concluded that: "*Financial incentives may be effective in changing healthcare professional practice. The evidence has serious methodological*

limitations and is also very limited in its completeness and generalisability. We found no evidence from reviews that examined the effect of financial incentives on patient outcomes." (Flodgren et al. 2011). A second review of reviews, over the same period, included 22 previous reviews as well as other recent papers that were "*not identified from an additional systematic review, but from our knowledge of the current evidence base on P4P effects"*. This seemed to have less stringent and some ad hoc inclusion criteria than the Flodgren review, and concluded with more negative conclusions (Eijkenaar, Emmert, Scheppach, and Schoffski 2013). Furthermore, our review is the most recent and anecdotally there seems to have been a proliferation of studies in the past 5 years with a large number using more robust study designs, such as difference-in-difference analysis.

Weaker study designs were more likely to show positive effects, suggesting that as study designs improve the likelihood of finding stronger effects will be lower, assuming nothing else changes. There seems to have been a general growth in the use of difference-in-difference designs. These designs are not currently included in Cochrane Collaboration study designs. However, they have emerged from the policy evaluation literature in economics, and have a number of advantages over other non-randomised designs, especially interrupted time series designs.

Schemes from the US had the same probability of finding an effect as non-US studies, and schemes focusing on hospitals had the same probability of finding an effect as studies targeting primary care, after controlling for how incentives were allowed to be used.

There are several notable findings in terms of the design of schemes. Overall, very few incentive design features that have been hypothesised to have an effect in theoretical models

and in the literature were statistically significant. This could be partly due to small sample sizes. In particular, the key innovation in the US has been the combination of rewards for pay for performance with rewards for reducing costs such as one and two-sided risk sharing models, yet this seems no better than pay for performance alone in terms of the proportion of positive outcomes. Many shared savings models are in their early stages, and so more evidence is required to examine if this persists over time.

A key finding is that schemes that reward for improvements in performance over time have a lower probability of being effective than those that do not. This is important to understand further as the dynamics of incentive schemes are complex. Schemes that did not reward for performance improvement included single threshold schemes but also other types of scheme such as value-based pricing of DRGs. The behavioural effects also depend on a range of more specific factors that could not be easily captured due to heterogeneity and small sample sizes, including the distance between measures/thresholds (i.e. the number of thresholds), whether payments are non-linear (e.g. increasing) at each time point/threshold, and whether the thresholds are set high or low in the distribution of performance.

No studies collected data on how providers used the incentive payments, rather they reported the rules. We find weak evidence that schemes allowing incentive funding to be used for specific (but non-physician income) purposes leads to a higher probability of an effect compared to physicians being allowed to use incentive funding as income. Again, more evidence is required as cell sizes were small. Since no study collected any data on exactly what the incentive funding was used for (all we could capture was rules), this effect could perhaps reflect the 'culture' of the scheme where quality improvements are more central than

financial concerns. Further studies should attempt to collect data on what incentive payments were used for, and report more clearly the rules of use.

The size of the incentives as a percentage of revenue was not associated with the probability of an effect. This is contrary to expectations. Though the sample size was small (22 schemes), the scatterplot did not show a clear relationship between incentive size and effect and so increasing the sample size may not make a difference if future studies are similar.

There are some limitations of the review. First, the review was not a systematic review in that; i) the search terms used were limited, ii) we did not conduct a full critical appraisal of the quality of the studies and their risk of bias. Nevertheless, given these limitations we believe we have identified the vast majority of empirical studies in the time period of interest. A major limitation is the use of vote counting rather than summarising effect sizes. Using vote counting gives an estimate of the probability that a scheme will be effective. However, vote counting gives equal weight to small but statistically significant effects that may not be 'clinically' or 'economically' significant, than it does to large effects. Vote counting ignores the quality of the evidence, where risk of bias can influence whether a scheme has a positive and statistically significant effect. However, when extracting data we made notes of any obvious risks of bias through unit of analysis error or small sample size. Though we classified study designs as to their quality and included studies which used designs that attempted to adjust for observed and unobserved confounding variables, these designs could have been poorly applied.

A further limitation is that there may a range of other factors that influence the probability of the scheme having an effect that were not possible to capture quantitatively. This includes

unobserved factors related to the how the scheme was developed, the extent to which clinicians were involved, and the extent of already existing quality improvement initiatives and public reporting. Though this is partly captured in the regression analysis using clustering, which accounts for unobserved factors within each scheme, within cluster sample sizes are on average small such that for 31 out of 44 schemes only one study was conducted. In addition, clustering (and random effects models) assumes the unobserved factors are uncorrelated with the independent variables; fixed effects models could not be implemented as there was not enough within cluster variation.

It was clear that many studies did not report some key basic design features, including the size of incentives relative to revenue, details about threshold payments, and other key aspects of incentive design. More consistent reporting of the characteristics of interventions is necessary. In addition, there are likely to be interactions between scheme design characteristics. For example, the size of the incentive may influence the effect of the other design features, such as the effect of paying for improvement. Larger sample sizes are required to test these hypotheses. As the use of financial incentives to encourage value-based health care grows, it remains important to conduct more research into the design of such schemes, and the impact of different designs on behaviour. There are still few randomised trials, but the use of difference in difference designs seems to be growing. Those implementing such schemes need to build rigorous evaluation into the implementation and roll out of schemes if knowledge is to improve.

References

Afendulis, C. C., A. M. Fendrick, Z. Song, B. E. Landon, D. G. Safran, R. E. Mechanic, and M. E. Chernew. 2014. The impact of global budgets on pharmaceutical spending and utilization: early experience from the alternative quality contract. *Inquiry* 51.

Allen, T., E. Fichera, and M. Sutton. 2016. Can Payers Use Prices to Improve Quality? Evidence from English Hospitals. *Health Economics* 25 (1):56-70.

Alshamsan, R., J. T. Lee, A. Majeed, G. Netuveli, and C. Millett. 2012. Effect of a UK pay-for-performance program on ethnic disparities in diabetes outcomes: interrupted time series analysis. *Ann Fam Med* 10 (3):228-34.

APM FPT Work Group. 2016. Alternative Payment Model Framework (APM). Final White Paper: Health Care Payment Learning and Action Network.

Arrowsmith, M. E., A. Majeed, J. T. Lee, and S. Saxena. 2014. Impact of pay for performance on prescribing of long-acting reversible contraception in primary care: an interrupted time series study. *PLoS One* 9 (4):e92205.

Bardach, N. S., J. J. Wang, S. F. De Leon, S. C. Shih, W. J. Boscardin, L. E. Goldman, and R. A. Dudley. 2013. Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: a randomized trial. *Journal of the American Medical Association* 310 (10):1051-9.

Berwick, Donald M, and Andrew D Hackbarth. 2012. Eliminating waste in US health care. *Journal of the American Medical Association* 307 (14):1513-1516.

Bhalla, R., C. B. Schechter, A. H. Strelnick, N. Deb, P. Meissner, and B. P. Currie. 2013. Pay for performance improves quality across demographic groups. *Qual Manag Health Care* 22 (3):199-209.

Burwell, Sylvia M. 2015. Setting Value-Based Payment Goals — HHS Efforts to Improve U.S. Health Care. *New England Journal of Medicine* 372 (10):897-899.

Bushman, BJ, and MC Wang. 1994. Vote-counting procedures in meta-analysis. *The Handbook of Research Synthesis* 236:193-213.

Cameron, A Colin, and Douglas L Miller. 2015. A practitioner's guide to cluster-robust inference. *Journal of Human Resources* 50 (2):317-372.

Chang, R. E., S. P. Lin, and D. C. Aron. 2012. A pay-for-performance program in Taiwan improved care for some diabetes patients, but doctors may have excluded sicker ones. *Health Aff (Millwood)* 31 (1):93-102.

Chen, J. Y., H. Tian, D. T. Juarez, I. Yermilov, R. S. Braithwaite, K. A. Hodges, A. Legorreta, and R. S. Chung. 2011. Does pay for performance improve cardiovascular care in a "real-world" setting? *Am J Med Qual* 26 (5):340-8.

Chen, P. C., Y. C. Lee, and R. N. Kuo. 2012. Differences in patient reports on the quality of care in a diabetes pay-for-performance program between 1 year enrolled and newly enrolled patients. *Int J Qual Health Care* 24 (2):189-96.

Chen, T. T., K. P. Chung, I. C. Lin, and M. S. Lai. 2011. The unintended consequence of diabetes mellitus pay-for-performance (P4P) program in Taiwan: are patients with more comorbidities or more severe conditions likely to be excluded from the P4P program? *Health Serv Res* 46 (1 Pt 1):47-60.

Cheng, AH, and JM Sutherland. 2013. British Columbia's pay-for-performance experiment: part of the solution to reduce emergency department crowding? *Health Policy* 113 (1-2):86-92.

Cheng, S. H., T. T. Lee, and C. C. Chen. 2012. A longitudinal examination of a pay-for-performance program for diabetes care: evidence from a natural experiment. *Med Care* 50 (2):109-16.

Chernew, M. E., R. E. Mechanic, B. E. Landon, and D. G. Safran. 2011. Private-payer innovation in Massachusetts: the 'alternative quality contract'. *Health Aff (Millwood)* 30 (1):51-61.

Chien, A. T., D. Eastman, Z. Li, and M. B. Rosenthal. 2012. Impact of a pay for performance program to improve diabetes care in the safety net. *Prev Med* 55 Suppl:S80-5.

Chien, A. T., Z. Li, and M. B. Rosenthal. 2010. Improving timely childhood immunizations through pay for performance in Medicaid-managed care. *Health Serv Res* 45 (6 Pt 2):1934-47.

Chien, A. T., Z. Song, M. E. Chernew, B. E. Landon, B. J. McNeil, D. G. Safran, and M. A. Schuster. 2014. Two-year impact of the alternative quality contract on pediatric health care quality and spending. *Pediatrics* 133 (1):96-104.

Colla, CH, and ES Fisher. 2014. Beyond PCMHs and accountable care organizations: payment reform that encourages customized care. *J Gen Intern Med* 29 (10):1325-7.

Colla, CH, PP Goodney, VA Lewis, BK Nallamothu, DJ Gottlieb, and E Meara. 2014. Implementation of a pilot accountable care organization payment model and the use of discretionary and nondiscretionary cardiovascular care. *Circulation* 130 (22):1954-61.

Colla, CH, VA Lewis, DJ Gottlieb, and ES Fisher. 2013. Cancer spending and accountable care organizations: Evidence from the Physician Group Practice Demonstration. *Healthc (Amst)* 1 (3-4):100-107.

Colla, CH, DE Wennberg, E Meara, JS Skinner, D Gottlieb, VA Lewis, CM Snyder, and ES Fisher. 2012. Spending differences associated with the Medicare Physician Group Practice Demonstration. *JAMA, The Journal of the American Medical Association* (10):1015.

Conrad, Douglas A., and Lisa Perry. 2009. Quality-Based Financial Incentives in Health Care: Can We Improve Quality by Paying for It? *Annual Review of Public Health* 30 (1):357-371.

Damberg, CL, ME Sorbero, SL Lovejoy, G Martsolf, L Raaen, and D Mandel. 2014. Measuring Success in Health Care Value-Based Purchasing Programs. Santa Monica: Rand Corporation.

de Walque, D., P. J. Gertler, S. Bautista-Arredondo, A. Kwan, C. Vermeersch, J. de Dieu Bizimana, A. Binagwaho, and J. Condo. 2015. Using provider performance incentives to increase HIV testing and counseling services in Rwanda. *J Health Econ* 40:1-9.

Department of Health. 2015. Evaluation report of the Diabetes Care Project. Canberra: Department of Health.

Eijkenaar, F, M Emmert, M Scheppach, and O Schoffski. 2013. Effects of pay for performance in health care: a systematic review of systematic reviews. *Health Policy* 110 (2-3):115-30.

Eijkenaar, Frank. 2013. Key issues in the design of pay for performance programs. *The European Journal of Health Economics* 14 (1):117-131.

Fiorentini, Gianluca, Elisa Iezzi, Matteo Lippi Bruni, and Cristina Ugolini. 2011. Incentives in Primary Care and Their Impact on Potentially Avoidable Hospital Admissions. *European Journal of Health Economics* 12 (4):297-309.

Flodgren, G, MP Eccles, S Shepperd, A Scott, E Parmelli, and FR. Beyer. 2011. An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes. In *Cochrane Database of Systematic Reviews*.

Gallagher, N., C. Cardwell, C. Hughes, and D. O'Reilly. 2015. Increase in the pharmacological management of Type 2 diabetes with pay-for-performance in primary care in the UK. *Diabet Med* 32 (1):62-8.

Gavagan, T. F., H. Du, B. G. Saver, G. J. Adams, D. M. Graham, R. McCray, and G. K. Goodrick. 2010. Effect of financial incentives on improvement in medical quality indicators for primary care. *J Am Board Fam Med* 23 (5):622-31.

Glasziou, Paul, Heather Buchan, Chris Del Mar, Jenny Doust, Mark Harris, Rosemary Knight, Anthony Scott, Ian A Scott, and Alexis Stockwell. 2012. When financial incentives do more good than harm: a checklist. *BMJ* 5047:345-350.

Hamilton, F. L., A. A. Laverty, D. Gluvajic, K. Huckvale, J. Car, A. Majeed, and C. Millett. 2014. Effect of financial incentives on delivery of alcohol screening and brief intervention (ASBI) in primary care: longitudinal study. *J Public Health (Oxf)* 36 (3):450-9.

Harrison, M. J., M. Dusheiko, M. Sutton, H. Gravelle, T. Doran, and M. Roland. 2014. Effect of a national primary care pay for performance scheme on emergency hospital admissions for ambulatory care sensitive conditions: controlled longitudinal study. *Bmj* 349:g6423.

Higgins, JPT, and S Green. 2011. *Cochrane handbook for systematic reviews of interventions*. Vol. Version 5.1.0.: The Cochrane Collaboration.

Houle, S. K., F. A. McAlister, C. A. Jackevicius, A. W. Chuck, and R. T. Tsuyuki. 2012. Does performance-based remuneration for individual health care practitioners affect patient care?: a systematic review. *Ann Intern Med* 157 (12):889-99.

Hsieh, H. M., S. L. Tsai, S. J. Shin, L. W. Mau, and H. C. Chiu. 2015. Cost-effectiveness of diabetes pay-for-performance incentive designs. *Med Care* 53 (2):106-15.

Jha, A. K., K. E. Joynt, E. J. Orav, and A. M. Epstein. 2012. The long-term effect of premier pay for performance on patient outcomes. *N Engl J Med* 366 (17):1606-15.

Jha, A. K., E. J. Orav, and A. M. Epstein. 2010. The effect of financial incentives on hospitals that serve poor patients. *Ann Intern Med* 153 (5):299-306.

Kawai, Alison Tse, Michael S. Calderwood, Robert Jin, Stephen B. Soumerai, Louise E. Vaz, Donald Goldmann, and Grace M. Lee. 2015. Impact of the Centers for Medicare and Medicaid Services Hospital-Acquired Conditions Policy on Billing Rates for 2 Targeted Healthcare-Associated Infections. *Infection Control And Hospital Epidemiology*:1-7.

Kiran, T., A. S. Wilton, R. Moineddin, L. Paszat, and R. H. Glazier. 2014. Effect of payment incentives on cancer screening in Ontario primary care. *Ann Fam Med* 12 (4):317-23.

Kontopantelis, E., T. Doran, H. Gravelle, R. Goudie, L. Siciliani, and M. Sutton. 2012. Family doctor responses to changes in incentives for influenza immunization under the U.K. Quality and Outcomes Framework pay-for-performance scheme. *Health Serv Res* 47 (3 Pt 1):1117-36.

Kontopantelis, E., D. Springate, D. Reeves, D. M. Ashcroft, J. M. Valderas, and T. Doran. 2014. Withdrawing performance indicators: retrospective analysis of general practice performance under UK Quality and Outcomes Framework. *Bmj* 348:g330.

Kristensen, S. R., R. Meacock, A. J. Turner, R. Boaden, R. McDonald, M. Roland, and M. Sutton. 2014. Long-term effect of hospital pay for performance on mortality in England. *N Engl J Med* 371 (6):540-8.

Kruse, G. B., D. Polsky, E. A. Stuart, and R. M. Werner. 2012. The impact of hospital pay-for-performance on hospital and Medicare costs. *Health Serv Res* 47 (6):2118-36.

Kruse, G. R., Y. Chang, J. H. Kelley, J. A. Linder, J. S. Einbinder, and N. A. Rigotti. 2013. Healthcare system effects of pay-for-performance for smoking status documentation. *Am J Manag Care* 19 (7):554-61.

Kuo, R. N., K. P. Chung, and M. S. Lai. 2011. Effect of the Pay-for-Performance Program for Breast Cancer Care in Taiwan. *J Oncol Pract* 7 (3 Suppl):e8s-e15s.

Lai, C. L., and Y. H. Hou. 2013. The association of clinical guideline adherence and pay-for-performance among patients with diabetes. *J Chin Med Assoc* 76 (2):102-7.

Langdown, C., and S. Peckham. 2014. The use of financial incentives to help improve health outcomes: is the quality and outcomes framework fit for purpose? A systematic review. *J Public Health (Oxf)* 36 (2):251-8.

Lee, Grace M, Ken Kleinman, Stephen B Soumerai, Alison Tse, David Cole, Scott K Fridkin, Teresa Horan, Richard Platt, Charlene Gay, and William Kassler. 2012. Effect of nonpayment for preventable infections in US hospitals. *New England Journal of Medicine* 367 (15):1428-1437.

Lee, J. T., G. Netuveli, A. Majeed, and C. Millett. 2011. The effects of pay for performance on disparities in stroke, hypertension, and coronary heart disease management: interrupted time series study. *PLoS One* 6 (12):e27236.

Lemak, C. H., T. A. Nahra, G. R. Cohen, N. D. Erb, M. L. Paustian, D. Share, and R. A. Hirth. 2015. Michigan's Fee-For-Value Physician Incentive Program Reduces Spending And Improves Quality In Primary Care. *Health Aff (Millwood)* 34 (4):645-52.

Lester, H., J. Schmittdiel, J. Selby, B. Fireman, S. Campbell, J. Lee, A. Whippy, and P. Madvig. 2010. The impact of removing financial incentives from clinical quality indicators: longitudinal analysis of four Kaiser Permanente indicators. *Bmj* 340:c1898.

Li, J., J. Hurley, P. DeCicca, and G. Buckley. 2014. Physician response to pay-for-performance: evidence from a natural experiment. *Health Econ* 23 (8):962-78.

Li, Y. H., W. C. Tsai, M. Khan, W. T. Yang, T. F. Lee, Y. C. Wu, and P. T. Kung. 2010. The effects of pay-for-performance on tuberculosis treatment in Taiwan. *Health Policy Plan* 25 (4):334-41.

Malhotra, A, D Maughan, J Ansell, R Lehman, A Henderson, M Gray, T Stephenson, and S Bailey. 2015. Choosing Wisely in the UK: the Academy of Medical Royal Colleges' initiative to reduce the harms of too much medicine. *Britha Medical Journal* 350 (h2308).

McDonald, R, S Zaidi, S Todd, F Konteh, K Hussain, J Roe, T Allen, E Fichera, and M Sutton. 2012. Qualitative and Quantitative Evaluation of the Introduction of Best Practice Tariffs. An evaluation report commissioned by the Department of Health: University of Nottingham.

McDonald, Ruth, Sabeeh Zaidi, Sarah Todd, Frederick Konteh, Kasser Hussain, Jim Roe, Thomas Allen, Eleonora Fichera, and Matthew Sutton. 2012. A qualitative and quantitative evaluation of the introduction of Best Practice Tariffs. *An evaluation report commissioned by the Department of Health. University of Nottingham.*

McKethan, A., and A. K. Jha. 2014. Designing smarter pay-for-performance programs. *Jama* 312 (24):2617-8.

McLeod, Hugh, Deirdre Blissett, Steven Wyatt, and Mohammed A. Mohammed. 2015. Effect of Pay-For-Outcomes and Encouraging New Providers on National Health Service Smoking Cessation Services in England: A Cluster Controlled Study. *PLoS ONE* 10 (4):e0123349.

McWilliams, J. M., M. E. Chernew, B. E. Landon, and A. L. Schwartz. 2015. Performance differences in year 1 of pioneer accountable care organizations. *N Engl J Med* 372 (20):1927-36.

McWilliams, J. M., B. E. Landon, and M. E. Chernew. 2013. Changes in health care spending and quality for Medicare beneficiaries associated with a commercial ACO contract. *Jama* 310 (8):829-36.

McWilliams, J. M., B. E. Landon, M. E. Chernew, and A. M. Zaslavsky. 2014. Changes in patients' experiences in Medicare Accountable Care Organizations. *N Engl J Med* 371 (18):1715-24.

Meacock, R., S. R. Kristensen, and M. Sutton. 2014. The cost-effectiveness of using financial incentives to improve provider quality: a framework and application. *Health Econ* 23 (1):1-13.

Nyweide, D. J., W. Lee, T. T. Cuerdon, H. H. Pham, M. Cox, R. Rajkumar, and P. H. Conway. 2015. Association of Pioneer Accountable Care Organizations vs Traditional Medicare Fee for Service With Spending, Utilization, and Patient Experience. *JAMA.*

O'Callaghan, Gerry, Hendrika Meyer, and Adam G Elshaug. 2015. Choosing wisely: the message, messenger and method. *The Medical journal of Australia* 202 (4):175-177.

Papke, Leslie E, and Jeffrey Wooldridge. 1993. Econometric methods for fractional response variables with an application to 401 (k) plan participation rates: National Bureau of Economic Research Cambridge, Mass., USA.

Peabody, John W., Riti Shimkhada, Stella Quimbo, Orville Solon, Xylee Javier, and Charles McCulloch. 2014. The Impact of Performance Incentives on Child Health Outcomes: Results from a Cluster Randomized Controlled Trial in the Philippines. *Health Policy and Planning* 29 (5):615-621.

Powell-Jackson, T., W. C. Yip, and W. Han. 2014. REALIGNING DEMAND AND SUPPLY SIDE INCENTIVES TO IMPROVE PRIMARY HEALTH CARE SEEKING IN RURAL CHINA. *Health Econ.*

Rajkumar, Rahul, Patrick H Conway, and Marilyn Tavenner. 2014. CMS—engaging multiple payers in payment reform. *Jama* 311 (19):1967-1968.

Rosenthal, MB, and RG Frank. 2006. What is the empirical basis for paying for quality in health care? *Medical Care Research and Review* 63:135 - 157.

Ryan, AM, and J Blustein. 2011. The effect of the MassHealth hospital pay-for-performance program on quality. *Health Serv Res* 46 (3):712-28.

Ryan, AM, J Blustein, and LP Casalino. 2012. Medicare's flagship test of pay-for-performance did not spur more rapid quality improvement among low-performing hospitals. *Health Aff (Millwood)* 31 (4):797-805.

Ryan, AM, JF Burgess, MF Pesko, WB Borden, and JB Dimick. 2015. The early effects of Medicare's mandatory hospital pay-for-performance program. *Health Serv Res* 50 (1):81-97.

Ryan, AM, CM McCullough, SC Shih, JJ Wang, MS Ryan, and LP Casalino. 2014. The intended and unintended consequences of quality improvement interventions for small practices in a community-based electronic health record implementation project. *Medical care* 52 (9):826-832.

Ryan, Andrew. 2009. Hospital-based pay-for-performance in the United States. *Health Economics* 18 (10):1109-1113.

Ryan, Andrew M, James F Burgess, and Justin B Dimick. 2015. Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences. *Health Services Research* 50 (4):1211-1235.

Saint-Lary, O., and J. Sicsic. 2015. Impact of a pay for performance programme on French GPs' consultation length. *Health Policy* 119 (4):417-26.

Scott, A, L Nacarella, J Furler, D Young, P Sivey, D Ouakrim, and L Willenberg. 2010. Using financial incentrives to improve the quality of primary care in Australia. Final Report. Canberra: Australian Primary Health Care Research Institute, Australian National University.

Scott, A, P Sivey, D Ait Ouakrim, L Willenberg, L Naccarella, J Furler, and D Young. 2011. The effect of financial incentives on the quality of health care provided by primary care physicians. *Cochrane Database of Systematic Reviews* Issue 4 (Art. No.: CD008451).

Scott, A., and M. F. Harris. 2012. Designing payments for GPs to improve the quality of diabetes care. *Med J Aust* 196 (1):24-6.

Serumaga, B., D. Ross-Degnan, A. J. Avery, R. A. Elliott, S. R. Majumdar, F. Zhang, and S. B. Soumerai. 2011. Effect of pay for performance on the management and outcomes of hypertension in the United Kingdom: interrupted time series study. *Bmj* 342:d108.

Sharp, A. L., Z. Song, D. G. Safran, M. E. Chernew, and A. Mark Fendrick. 2013. The effect of bundled payment on emergency department use: alternative quality contract effects after year one. *Acad Emerg Med* 20 (9):961-4.

Shih, T., L. H. Nicholas, J. R. Thumma, J. D. Birkmeyer, and J. B. Dimick. 2014. Does pay-for-performance improve surgical outcomes? An evaluation of phase 2 of the Premier Hospital Quality Incentive Demonstration. *Ann Surg* 259 (4):677-81.

Simpson, C. R., P. C. Hannaford, L. D. Ritchie, A. Sheikh, and D. Williams. 2011. Impact of the pay-for-performance contract and the management of hypertension in Scottish primary care: a 6-year population-based repeated cross-sectional study. *Br J Gen Pract* 61 (588):e443-51.

Song, Z. 2014. Accountable Care Organizations in the U.S. Health Care System. *J Clin Outcomes Manag* 21 (8):364-371.

Song, Z., A. M. Fendrick, D. G. Safran, B. Landon, and M. E. Chernew. 2013. Global Budgets and Technology-Intensive Medical Services. *Healthc (Amst)* 1 (1-2):15-21.

Song, Z., S. Rose, D. G. Safran, B. E. Landon, M. P. Day, and M. E. Chernew. 2014. Changes in health care spending and quality 4 years into global payment. *N Engl J Med* 371 (18):1704-14.

Song, Z., D. G. Safran, B. E. Landon, Y. He, R. P. Ellis, R. E. Mechanic, M. P. Day, and M. E. Chernew. 2011. Health care spending and quality in year 1 of the alternative quality contract. *N Engl J Med* 365 (10):909-18.

Song, Z., D. G. Safran, B. E. Landon, M. B. Landrum, Y. He, R. E. Mechanic, M. P. Day, and M. E. Chernew. 2012. The 'Alternative Quality Contract,' based on a global budget, lowered medical spending and improved quality. *Health Aff (Millwood)* 31 (8):1885-94.

Sun, Xiaojie, Xiaoyun Liu, Qiang Sun, Winnie Yip, Adam Wagstaff, and Qingyue Meng. 2014. The impact of a pay-for-performance scheme on prescription quality in rural China : an impact evaluation. The World Bank, Policy Research Working Paper Series: 6892. Washington: The World Bank, Policy Research Working Paper Series: 6892.

Sutton, M., S. Nikolova, R. Boaden, H. Lester, R. McDonald, and M. Roland. 2012. Reduced mortality with hospital pay for performance in England. *N Engl J Med* 367 (19):1821-8.

Swaminathan, S., V. Mor, R. Mehrotra, and A. N. Trivedi. 2014. Effect of Medicare Dialysis Payment Reform on Use of Erythropoiesis Stimulating Agents. *Health Serv Res*.

Tan, E. C., R. F. Pwu, D. R. Chen, and M. C. Yang. 2014. Is a diabetes pay-for-performance program cost-effective under the National Health Insurance in Taiwan? *Qual Life Res* 23 (2):687-96.

Tsai, W. C., P. T. Kung, M. Khan, C. Campbell, W. T. Yang, T. F. Lee, and Y. H. Li. 2010. Effects of pay-for-performance system on tuberculosis default cases control and treatment in Taiwan. *J Infect* 61 (3):235-43.

Unutzer, J., Y. F. Chan, E. Hafer, J. Knaster, A. Shields, D. Powers, and R. C. Veith. 2012. Quality improvement with pay-for-performance incentives in integrated behavioral health care. *Am J Public Health* 102 (6):e41-5.

Van Herck, P., D. De Smedt, L. Annemans, R. Remmen, M. B. Rosenthal, and W. Sermeus. 2010. Systematic review: Effects, design choices, and context of pay-for-performance in health care. *BMC Health Serv Res* 10:247.

Witter, S., A. Fretheim, F. L. Kessy, and A. K. Lindahl. 2012. Paying for performance to improve the delivery of health interventions in low- and middle-income countries. *Cochrane Database Syst Rev* 2:Cd007899.

Yip, W., T. Powell-Jackson, W. Chen, M. Hu, E. Fe, M. Hu, W. Jian, M. Lu, W. Han, and W. C. Hsiao. 2014. Capitation combined with pay-for-performance improves antibiotic prescribing practices in rural China. *Health Aff (Millwood)* 33 (3):502-10.

Yu, H. C., W. C. Tsai, and P. T. Kung. 2014. Does the pay-for-performance programme reduce the emergency department visits for hypoglycaemia in type 2 diabetic patients? *Health Policy Plan* 29 (6):732-41.

Figure 2. Relationship between size of incentives and proportion of positive outcomes (for 22/44 schemes reporting size)

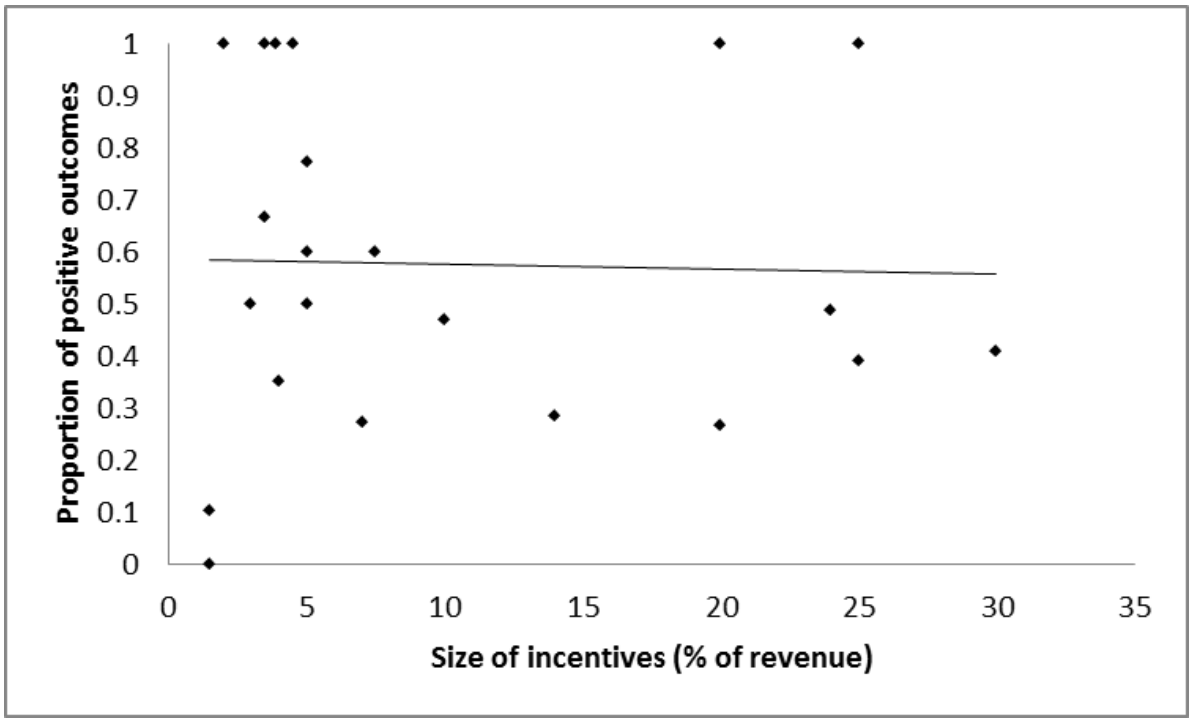

Table 1. Number of studies and schemes included in the review, by country and setting

	Hospitals (# of studies)	Multi-specialty/primary care physicians (# of studies)	Number of studies	Number of schemes
United States	14	28	42	25
United Kingdom	5	11	16	5
Taiwan	3	6	9	3
Canada	1	2	3	2
China	0	3	3	2
Italy	1	1	2	2
Australia	0	2	2	2
France	0	1	1	1
Philippines	1	0	1	1
Rwanda	0	1	1	1
Total	**25**	**55**	**80**	**44**

Table 2. Summary of overall effects.

	Total number	**Number positive and statistically significant**	**Proportion**
Outcome measures	1304	595	0.46
By incentive scheme (n=44):	**Mean**	**Median**	**Min-max**
Number of positive outcomes per scheme	13.52	4	0 to 156
Number of outcomes per scheme	29.64	7.5	1 to 332
Proportion of positive outcomes per scheme	0.56		0 to 1
By empirical study (n=80):			
Number of positive outcomes per study	7.43	3.5	0 to 68
Number of outcomes per study	16.30	9	1 to 120
Proportion of positive outcomes per study	0.54		0 to 1

Table 3. Ranking of schemes by the mean proportion of positive outcomes per scheme

	Proportion of positive outcomes	Number of positive outcomes	Number of outcomes	Number of studies
Breast Cancer P4P	1.000	3	3	1
ED P4P	1.000	2	2	1
Health Plan in Hawaii	1.000	4	4	1
Hudson Health Plan - immunisation	1.000	2	2	1
Kaiser Permanente Northern California	1.000	2	2	1
Lazio DRG P4P	1.000	4	4	1
Medicare Dialysis payment Reform	1.000	1	1	1
Mental health integration program (MHIP)	1.000	1	1	1
P4P for Tuberculosis	1.000	2	2	2
Partners HealthCare, Inc.	1.000	1	1	1
RWF-AHRG Health Promotion	1.000	1	1	1
Clinical directed enhanced services	0.833	5	6	1
BSBC (Michigan) Physician Group Incentive Program	0.773	17	22	1
P4P for Diabetes	0.723	47	65	6
Houston/Harris County Community Health Program	0.667	2	3	1
Physician Group Practice Demonstration	0.623	76	122	3
Bronx CHAMPION	0.600	3	5	1
Primary Care Information Project	0.600	12	20	2
Veterans Affairs networks	0.583	7	12	1
Medicare Never Events	0.571	4	7	2
Medicare Pioneer ACOs	0.505	55	109	3
Emilia-Romagna P4P	0.500	1	2	1
Ontario P4P	0.500	4	8	2
Practice Incentives Program	0.500	2	4	1
Quality improvement demonstration study (QIDS)	0.500	2	4	1
Best Practice Tariff	0.488	20	41	2
Alternative Quality Contract	0.470	156	332	8
Ningxia scheme	0.408	20	49	2
Quality and Outcomes Framework	0.389	44	113	9
eHeart (pilot)	0.381	8	21	1
Advancing Quality	0.350	21	60	3
Alabama Managed Care Organisation	0.333	3	9	1
NHS stop smoking services	0.333	2	6	1
Diabetes Care Project	0.324	11	34	1
Fairview Health Services	0.308	12	39	1
Medicare Advantage Prescription Drug Plan (MAPD)	0.294	5	17	1
Rwanda P4P	0.286	2	7	1
PacificCare / IHA	0.278	15	54	1

Contract for Improving Individual Practices' (CAPI)	0.273	3	11	1
Shandong scheme	0.267	8	30	1
Medicare Premier Hospital Quality Incentive Demonstration	0.102	5	49	5
Hudson Health Plan - diabetes	0.000	0	12	1
Medicaid MassHealth	0.000	0	2	1
Medicare Hospital Value-Based Purchasing (HVBP)	0.000	0	6	1

Table 4. Ranking of countries by the mean proportion of positive outcomes per scheme

	Proportion of positive outcomes per scheme (mean per country)	Number of outcomes	Number of studies	Number of schemes
Taiwan	0.91	70	9	3
Canada	0.75	10	3	2
Italy	0.75	6	2	2
US	0.56	853	42	25
Philippines	0.50	4	1	1
UK	0.48	226	16	5
Australia	0.41	38	2	2
China	0.34	79	3	2
Rwanda	0.29	7	1	1
France	0.27	11	1	1
All Non-US	0.56	451	38	19

Table 5. Factors associated with the proportion of positive outcomes per study[1].

	Model 1		Model 2	
	Marginal effect (se)	Sample mean	Marginal effect (se)	Sample mean
DID[2]	-0.246 (0.089)**	0.442	-0.190 (0.085)**	0.425
ITS[2]	-0.207 (0.104)**	0.156	-0.132 (0.116)	0.15
RCT[2]	-0.252 (0.098)**	0.104	-0.224 (0.106)**	0.1
US[3]	-0.083 (0.071)	0.519	-0.080 (0.067)	0.525
Hospital[4]	-0.060 (0.080)	0.325	-0.069 (0.076)	0.313
Costs and quality (APM3&4)[5]	0.107 (0.080)	0.364	0.099 (0.075)	0.35
Reward for improvement[6]	-0.224 (0.065)**	0.455	-0.206 (0.063)***	0.438
Discretionary use[7]	0.056 (0.089)	0.649	-	-
Specific use[7]	0.240 (0.143)*	0.065	-	-
n	77		80	
Psuedo-logl	-38.28		-40.55	
Clusters	41		44	
Obs per cluster (min-max)	1.9 (1-9)		1.8 (1-9)	
AIC	1.25		1.21	
BIC	-253		-275	

Notes: 1: *** p-value<0.0001; ** 0.0001<p-value<=0.05; * 0.05<p-value<0.1. 2: Omitted reference group is before and after designs, controlled before and after, and case control studies. 3: Omitted reference group is all other countries. 4: Omitted reference group is primary care. 5: Omitted reference group is pay for performance only (APM Category 2). 6: Omitted reference group is not rewarding for improvement.7: Omitted reference group is using incentives for physician income.

Table 6. Percent of revenue affected by the incentive scheme, by scheme[1].

Scheme	Country	Size of payment relative to revenue
Medicare Hospital Value-Based Purchasing (HVBP) (Ryan, Burgess, Pesko, Borden, and Dimick 2015)	US	1% to 2%
Medicare Premier Hospital Quality Incentive Demonstration (Jha, Joynt, Orav, and Epstein 2012; Shih, Nicholas, Thumma, Birkmeyer, and Dimick 2014)	US	1% to 2%
Medicare Dialysis Payment Reform (Swaminathan, Mor, Mehrotra, and Trivedi 2014)	US	2%
Health Plan in Hawaii (Chen et al. 2011)	US	3.5%
Houston/Harris County Community Health Program (Gavagan et al. 2010)	US	3% to 4%
Partners Health Care, Inc. (Kruse et al. 2013)	US	3% to 4.8%
Advancing Quality Initiative (Kristensen et al. 2014; Sutton et al. 2012)	UK	4%
Bronx CHAMPION (Bhalla et al. 2013)	US	5%
Primary Care Information Project (Bardach et al. 2013; Ryan et al. 2014)	US	5%
Quality improvement demonstration study (QIDS) (Peabody et al. 2014)	Philippines	5%
Breast Cancer P4P (Kuo, Chung, and Lai 2011)	Taiwan	2% to 7%
Contract for Improving Individual Practices (Saint-Lary and Sicsic 2015)	France	7%
Ontario P4P (Kiran, Wilton, Moineddin, Paszat, and Glazier 2014; Li, Hurley, DeCicca, and Buckley 2014)	Canada	3%-10%
Blue Cross Blue Shield of Michigan's Physician Group Incentive Program (Lemak et al. 2015)	US	5%-10%
Alternative Quality Contract (Sharp, Song, Safran, Chernew, and Mark Fendrick 2013; Song, Fendrick, Safran, Landon, and Chernew 2013; Song et al. 2011)	US	10%
Rwanda P4P (de Walque et al. 2015)	Rwanda	14%
Hudson Medicaid Health Plan (Chien, Li, and Rosenthal 2010)	US	15% to 25%
Best Practice Tariff (Allen, Fichera, and Sutton 2016; McDonald et al. 2012)	UK	24%
Mental Health Integration Program (Unutzer et al. 2012)	US	25%
Quality and Outcomes Framework (Lee, Netuveli, Majeed, and Millett 2011; Serumaga et al. 2011; Simpson, Hannaford, Ritchie, Sheikh, and Williams 2011)	UK	25%
Shandong scheme (Sun et al. 2014)	China	up to 20% max
Ningxia scheme (Powell-Jackson, Yip, and Han 2014; Yip et al. 2014)	China	up to 30% max

Notes: 1. The references include those where the size of the payment was mentioned.2. The regression in Table 5 used mid-points of ranges as independent variables where size is presented as a range.

Figure 1. Number of papers reviewed and included.