**The Effectiveness of Payment for Performance in health care: a meta-analysis and exploration of variation in outcomes**

Yewande K Ogundeji\* (PhD),1 J Martin Bland (PhD)2 and Trevor A Sheldon (DSc)3

1 Department of Health Sciences, University of York, York, UK YO10 5DD; Health Strategy and Delivery Foundation (HSDF), 1980 Wikki Spring Street, Maitama, Abuja, Nigeria

2 Department of Health Sciences, University of York, York, UK YO10 5DD

3 Hull York Medical School, University of York, York, UK YO10 5DD

**Corresponding author**

Dr Yewande Ogundeji

Address: Health Strategy and Delivery Foundation (HSDF), 1980 Wikki Spring Street, Maitama, Abuja, Nigeria

Phone number: +2348172289973

Email: ykogundeji@gmail.com

The authors do not have any conflicting interests to declare

**The Effectiveness of Payment for Performance in health care: a meta-analysis and exploration of variation in outcomes**

**Abstract**

**Background**

Pay for performance (P4P) incentive schemes are increasingly used world-wide to improve health system performance but results of evaluations vary considerably. A systematic analysis of this variation in the effects of P4P schemes is needed.

**Methods**

Evaluations of P4P schemes from any country were identified by searching for and updating systematic reviews of P4P schemes in health care in four bibliographic databases. Outcomes using different measures of effect were converted into standardised effect sizes and each study was categorised as to whether or not it found a positive effect. Subgroup analysis, meta-regression and multilevel logistic regression were used to investigate factors explaining heterogeneity. Random-effects models were used because they take into account heterogeneity likely to be due to differences between studies rather than just chance. Sensitivity analysis was used to test the effect of different assumptions.

**Findings**

96 primary studies were identified; 37 were included in the meta-analysis and meta-regression and all 96 in the logistic regression. The proportion of observed variation in study results that can be explained by true heterogeneity (I2) was 99.9%. Estimates of effect of P4P schemes were lower in evaluations using randomised controlled trials (SMD=0∙08; 95% CI: 0∙01 to 0∙15) compared to no controls (0∙15; 95%CI: 0∙09 to 0∙21), and lower for those measuring outcomes (e.g. smoking cessation) (SMD=0∙0; 95%CI: -0∙01 to 0∙01) compared to process measures (e.g. giving cessation advice) (0∙18; 95%CI: 0∙06 to 0∙31).

Adjusting for other design features and the evaluation method, the odds of showing a positive effect was three times higher for schemes with larger incentives (>5% of salary/usual budget) (OR = 3∙38; 95%CI: 1∙07 to 10∙64). There were non-statistically significant increases in the odds of success if the incentive is paid to individuals (as opposed to groups) (OR= 2∙0; 95%CI: 0∙62 to 6∙56) and if there is a lower perceived risk of not earning the incentive (OR= 2∙9; 95%CI: 0∙78 to 10∙83). Schemes evaluated using less rigorous designs were 24 times more likely to have positive estimates of effect than those using randomised controlled trials (OR = 24; 95%CI: 6∙3 to 92∙8).

**Interpretation**

Estimates of the effectiveness of incentive schemes on health outcomes are probably inflated due to poorly designed evaluations and a focus on process measures rather than health outcomes. Larger incentives and reducing the perceived risk of non-payment may increase the effect of these schemes on provider behaviour.

**Introduction**

Performance-based financing of health care or pay for performance (P4P) is increasingly used around the world as a mechanism to improve health system performance. Through the use of incentives linked to the achievement of metrics or targets it is hoped to improve delivery, utilisation, efficiency or outcomes of health care or pubic health services. There have been many evaluations of these schemes in different countries and several reviews of these studies.[1-3](#_ENREF_1) These reviews show that the evidence regarding its effectiveness is inconclusive and of limited use in informing policy due to the large variation in results of the evaluations.[4](#_ENREF_4),[5](#_ENREF_5) Heterogeneous results observed across P4P schemes might be explained by variation in design features, contexts, implementation factors, and evaluation design between the schemes.[1](#_ENREF_1),[6](#_ENREF_6) There are, however, no studies that explore heterogeneity in a structured quantitative way.

Given the increasing popularity of such schemes and their cost implications, it is important to analyse the results of these evaluations in more detail in order to explore the extent to which patterns exist which may have policy and practice significance. This paper systematically explores the extent and sources of heterogeneity in the results of evaluations of P4P schemes to identify features associated with success in P4P schemes.

**Methods**

We conducted a systematic review and meta-analysis.

*Literature search*

Evaluated P4P schemes were identified from published reviews of evaluations of the effectiveness of P4P. In addition, we updated one of the best systematic reviews,[1](#_ENREF_1) which scored 11/11 on the AMSTAR checklist, conducting the search up until April 2016 (Supplementary files S1-S5).[7](#_ENREF_7) Electronic database searches for systematic reviews were conducted in Database of Abstracts and Reviews of Effect (DARE), National Health Service Economic Evaluation Database (NHS EED), Health Technology Assessment (HTA)], Cochrane, and PubMed using the following keywords: financial incentives, performance based financing, and pay for performance. Websites and databases of health organisations involved in implementing and evaluating P4P were also searched e.g. The World Bank, Global Alliance for Vaccines and Immunisations (GAVI), and Cordaid.

There were no date or language restrictions. We included only primary studies that evaluated the impact of P4P on health service provider performance or quality of care.

*Potential sources of heterogeneity*

A template was used for data extraction to include: country of implementation, sample size and raw numbers of events (see supplementary file S6), the domain of performance (whether or not processes or outcomes measures were incentivised). We recorded three key design features of each evaluated scheme using a newly developed and validated P4P typology by Ogundeji *et al*: who receives the incentives (individuals or groups), size of incentive (large or small), and perceived risk of not earning the incentive (low or high) (see supplementary file S7).[8](#_ENREF_8) We also categorised the design of each evaluation to indicate whether it was a randomised controlled trial (RCT), a quasi-experimental study such as interrupted time series (statistical testing for a change in the outcome rate in measurements taken at ordered time periods before and after intervention) or before and after studies (as less well controlled studies can be more susceptible to bias).[9-12](#_ENREF_9)

*Statistical analysis*

1. Creating comparable measures of effect between studies

The measures of effect reported in the primary studies included: odds ratios, percentage point differences, means, and mean differences[[1]](#footnote-1). Therefore, we converted them into two common measures which could be compared across studies and combined.

First we converted them to standardised mean differences (and associated standard errors). This could only be done where data on absolute differences (percentages or numbers), sample size, standard deviations or standard errors or variance were reported.[13](#_ENREF_13),[14](#_ENREF_14) Some primary studies reported multiple principal outcome measures, for example, prescribing conduct, smoking cessation, and blood pressure reduction. If these were all included in the analyses without appropriate handling, it would overestimate the amount of independent information, so producing overly precise and possibly biased estimates.[15](#_ENREF_15),[16](#_ENREF_16) Selecting a primary outcome measures from the multiple outcomes reported was difficult, as the indicators/measures incentivised and reported covered different clinical areas. In these cases we computed a summary effect and its associated standard error using the formulae suggested by Borenstein et.al.[14](#_ENREF_14) (See supplementary file S8). We cautiously assumed a correlation of 0∙5 between outcomes in different clinical areas (e.g. smoking cessation and hospital mortality) and 0∙75 for outcomes in similar clinical areas (e.g. cholesterol and blood pressure levels in diabetic patients).[14](#_ENREF_14),[17](#_ENREF_17) We also conducted a sensitivity analysis using lower correlation values of 0∙5 and 0∙25 respectively.

A second approach to dealing with multiple outcome measures was to convert measures of effect to binary outcomes, coded according to whether or not the evaluation found that the P4P was effective. This approach requires less consistent data reporting and so increases the number of studies included, though losing information. We defined effectiveness as a statistically significant (P<0∙05) difference favoring the use of P4P over control groups. Because failure to find a statistical significance might be due to low power (small sample size), we also conducted a sensitivity analysis classifying outcomes with statistically non-significant positive effects as effective.

1. Extent of heterogeneity

An ‘inverse-variance weighted DerSimonain-Laird random-effects model’ was used to conduct the meta-analysis using Stata statistical package (version 12) ‘metan’ command. A random effects model was used because the observed differences among study results are likely to be due to differences between the studies and not chance.[14](#_ENREF_14),[18](#_ENREF_18) We pooled all relevant included studies and quantified heterogeneity using the I2 test statistic, which describes the percentage of the variability in effect estimates that is due to heterogeneity rather than chance.[19](#_ENREF_19) Publication bias was assessed by a visual inspection of a funnel plot.

1. Exploring heterogeneity

We undertook three different analyses to explore the potential sources of heterogeneity: sub-group analysis, meta-regression and multilevel logistic regression.

We stratified the meta-analysis into subgroups by (i) evaluation design (presenting separate pooled estimates of RCTs, quasi-experimental studies, and studies with no adequate control group) and (ii) domain of performance (process, intermediate outcomes, and health outcomes). In addition, we extended the subgroup analysis to reflect the different categories of quasi-experimental studies. This included pre-post design with control groups, cross sectional, interrupted time series (ITS), and other longitudinal studies.

A few studies that reported aggregate results from the analysis of mostly process measures with a few intermediate outcome measures were classified as studies using process measures. This is because in most cases, process measures have been shown to have an influence on related intermediate outcome measures.[8](#_ENREF_8)

We used meta-regression and a multilevel logistic regression model to explore the effects of the three design features (who receives the incentive, size of incentive, and risk of not earning the incentive) and the adequacy of control groups in the evaluation. Two sets of analyses were performed in each model; a univariate analysis for each explanatory variable and second, a multivariate analysis with each set of explanatory variables in the model simultaneously (Table 1).

For the meta-regression model, where available we used standardized mean difference (SMD) and 95% confidence intervals. The multilevel logistic regression model used the binary outcome of statistically significant effect or no effect of the P4P scheme.

Meta-regression analysis was performed in Stata 12 using the ‘metareg’ command which performs random-effects regression using aggregate level data. This extends the variance-weighted least squares regression by estimating an extra additive component of variance, which requires specification of the standard errors within each study.[20](#_ENREF_20) The random effects model considers the residual heterogeneity among intervention effects not modeled by the explanatory variables.[21](#_ENREF_21)

Several schemes had more than one evaluation, and within each evaluation there were sometimes more than one outcome. This nested or clustered structure (outcomes within studies within schemes) led us to use a multilevel logistic regression analysis which was performed in Stata 12 using the ‘xtmelogit’ command. This analyses clustered data by accounting for intra-class correlation (by correcting standard errors) and correcting denominator degrees of freedom for number of clusters (giving an estimate of the variability caused by each level of cluster).[22](#_ENREF_22) The fixed effect part of the model presents estimates that are interpreted similarly to a logistic regression (with robust standard errors) and the random effects part of the model accounts for the clustering effect or the amount of unexplained variation at each level.[16](#_ENREF_16)

**Results**

**Study characteristics**

Through the reviews and the updated search, 96 P4P evaluation studies (Figure 1) evaluating 68 different incentive programmes which met our inclusion criteria were identified. Included and excluded studies are shown in supplementary files S9 and S10. The characteristics of the included studies are summarised in supplementary file S11. Schemes had different scopes and sizes, and were implemented in different countries and contexts (ownership and funding), with varying levels of complexity.

There were 270 outcomes (from the 96 studies) across different areas of health care, such as cancer or tuberculosis screening, curative care, disease management, patient satisfaction, smoking cessation advice, hospital mortality, vaccinations, and utilization of health services.

All 96 studies could be included in the logistic regression and 37 reported sufficient data to be included in the meta-analysis and meta-regression (see supplementary files S12 and S13).

*Meta-analysis*

Figure 2 shows a forest plot of the 37 studies[23-59](#_ENREF_23) included in the meta-analysis, stratified by evaluation design (RCT=6, quasi-experimental=11, and no control groups/before and after studies=20). Figure 3 shows them by domain of performances (processes or outcomes).

The pooled overall effect is SMD = 0∙13, (95% CI: 0∙02 to 0∙24, p<0∙0001), which indicates a small but statistically significant overall positive effect of P4P. There was considerable heterogeneity I2 = 99∙9% (CI=99∙91 to 99∙92).[19](#_ENREF_19) The funnel plot is roughly symmetrical (see supplementary file S14). However, the Begg and Mazumdar test did indicate evidence of asymmetry (tau = 0.31, P = 0.008). Inspection of the graph suggests that the two points at the top right of the plot, papers by Kuo et al., 2011 and Lee et al., 2010, both from Taiwan, may be responsible for this. If we remove these two studies, the test is no longer significant (tau = 0.13, P = 0.3).

The subgroup analysis by study design (Figure 2) found that evaluations using RCTs had a much smaller overall positive effect (SMD= 0∙08, 95% CI: 0∙01 to 0∙15, I2= 0∙0% CI: 0∙00 to 74∙62) compared to the quasi experimental studies (0∙12, 95% CI: -0∙05 to 0∙29; I2= 99∙90% CI: 99∙89 to 99∙90), and those with no control group (0∙15, 95%CI: 0∙09 to 0∙21; I2= 96∙06% CI: 94∙31 to 97∙27). More granular analysis of categories of quasi-experimental studies (supplementary file S15) showed that evaluations using either interrupted time series designs or cross sectional studies had lower estimates of effect (SMD= 0∙07, 95% CI: -0∙03 to 0∙17; I2 = 73.52% CI: 11∙33 to 92∙10) and (SMD= 0∙03, 95% CI: 0∙01 to 0∙05; I2 = 0.00% CI: 00∙00 to 64∙80) respectively, compared to pre-post studies with control group (SMD= 0∙22, 95% CI: -0∙03 to 0∙47; I2 = 99∙96% CI: 99∙95 to 99∙96). Studies under the pre-post with control group appeared to have a larger effect estimate compared to the other quasi-experimental groups and studies with no controls (SMD= 0∙22, 95% CI: -0∙03 to 0∙47). This is could be because the pooled estimate is being positively skewed by two apparent outliers (Kuo et al., 2011 and Lee et al., 2010, both set in Taiwan) with large study sizes and large effect sizes.

Stratification by the domain of performance measured (Figure 3) found that schemes measuring success by improvements in *outcomes* such as reduction in hospital mortality or smoking cessation showed no effect (SMD=0∙00, 95%CI: -0∙01 to 0∙01) whereas those using *intermediate outcomes* such as blood pressure or cholesterol reduction (0∙07 95%CI, -0∙01, 0∙15) or *process measures* such as cancer screening or smoking cessation advice (0∙18, 95%CI: 0∙06 to 0∙31) showed larger effects.

*Meta-regression*

Table 2 shows the output from the univariate and multivariate analyses using standardised mean difference (SMD). These indicate that larger incentives and low risk of not earning the incentive increases the size of the estimate of the effect. However, all of these showed very small effects[60](#_ENREF_60) and none of these were statistically significant (I2=98.4%). However, given that I2 measures may be inflated or deflated by high or low precision of included studies[61](#_ENREF_61), another possible reason that I2 estimates are high could be because of the number of included studies with small standard errors (high precision).

*Multilevel logistic regression*

A likelihood ratio test of the multilevel logistic regression against an ordinary logistic regression model indicates that the multilevel model is a better fit to the data (Chi2 (df=2) 24∙43 P< 0∙0001). Table 3 shows that P4P schemes using large incentives had over three times the odds of reporting a positive effect. Similarly schemes where there is a lower perceived risk of not getting the incentive payment were nearly three times more likely to show a positive effect. Incentives paid to individuals rather than groups may be more effective, but the result was not statistically significant. As shown before, poor evaluation design was associated with a large increase in the odds of showing a positive effect (OR = 24; 95%CI: 6∙3 to 93).

The random effects parameters of the multilevel logistic regression model are shown in supplementary file S16; this estimates variation that is unaccounted for through the standard deviations at each level. There is no evidence of unexplained variation at the scheme/programme level, but there is evidence of some unexplained variation at the study level. A standard deviation of 1.83 at the study level indicates that outcomes in an evaluated P4P study which is one standard deviation above the mean have odds of being effective 6.2 times higher than comparable outcomes in an average evaluated P4P study [exp (1.83) = 6.20]

*Sensitivity analyses*

Results of sensitivity analyses performed in both regression models showed no material changes or lead to different conclusions compared to the original models (see supplementary file S17-S18).

**Discussion**

To our knowledge this is the first systematic quantitative exploration showing how design features and method of evaluation may influence the effectiveness of health care incentive schemes.

Overall, we found that although 70% of the outcome variables measured showed a positive effect, the overall size of the effects of P4P schemes was very modest (standardised effect size of only 0∙14). The impact of such schemes on health outcomes is probably even lower; studies using final outcomes as the basis for payment (such as mortality, morbidity or patient behaviour change) had a much lower effect than those using only process or intermediate outcome measures as the basis of payment (such as the supply of services or their uptake). Even this is likely to be an overestimate as we found that schemes evaluated using randomised controlled trials and other rigorous designs showed little or no effect.

We found that, in line with theory, larger incentives and reducing the risk of non-payment increases the likelihood of a positive effect and the size of that effect. Payments to individuals are more effective than to groups, but this was not statistically significant.

*Limitations*

Due to inadequate reporting of effect estimates, study sample sizes and standard errors, 60 out of 96 studies could not be included in the meta-analysis and the meta-regression. This led to loss of information and statistical power to detect the small effect sizes as statistically significant, and limited the number of variables that could be explored simultaneously.[14](#_ENREF_14) We mitigated this partially by converting the outcomes variables of all 96 studies to a binary variable (whether or not there was a statistically significant positive effect), thus allowing inclusion of all P4P evaluation studies in a multilevel logistic regression model. However, whilst this increases the number of studies, the analysis has low power to detect anything but large associations as statistically significant. It also loses information about the size of the effect; studies with small and large positive effects are treated equally. Using the criterion of a statistically significant positive effect is a rather crude way to categorise studies, however, reclassifying the studies so that any positive effect (whether or not statistically significant) was categorised as positive did not materially alter the results.

The categorisation of P4P schemes as incentivising process change when in fact they use a mixture of mainly process but also some outcome measures introduces some error and might over-estimate the impact of such schemes.

Large heterogeneity estimates observed (using the I2) could have been exaggerated as a result of the large number of studies.[61](#_ENREF_61) However, there was a high likelihood of underlying heterogeneity due to factors of policy and methodological significance which further justified the decision to explore variation in a systematic way.[62](#_ENREF_62)

A major potential source of bias in meta-analyses is the selective publication of positive results more likely in smaller studies. However, our funnel plot shows no evidence of this, a key strength in this analysis. Although, the Begg and Mazumdar test did indicate evidence of asymmetry, likely as a result of 2 studies from Taiwan by Kuo et al., 2011 and Lee et al., 2010, which nullifies the significant value of the test when excluded.

The size of incentive was the most influential design feature in both models. Schemes paying incentives of >5% of clinician’s salary or hospital/group budget) are more effective compared to those paying less. This suggests that incentive size is one of the most important design features to consider for the effectiveness of P4P. This confirms evidence from other studies.[63](#_ENREF_63),[64](#_ENREF_64) However, theoretical literature suggesting that clinicians often aspire to reach a ‘target income’[65](#_ENREF_65),[66](#_ENREF_66) would imply that clinicians close to their target income, may not be as responsive; in such cases the incentive may not stimulate desired behaviours.

The more money spent on incentives to improve performance, however, the less likely it is that the scheme will be cost effective. Few evaluations consider whether the extra costs of P4P schemes are justified by the additional health benefits derived from the improvements performance[67](#_ENREF_67) or indeed whether similar improvements could have been more effectively and efficiently achieved by other means.[68](#_ENREF_68)

The importance of the perceived risk of not earning the incentive is consistent with results from the review of reviews by Eijkenaar.[69](#_ENREF_69) This risk is smaller if the health care provider is more confident that they can make the necessary changes and reach any targets set and that if they make the necessary changes, they will actually get paid. Thus low risk is indicated by: a short time lag between verification of performance and receipt of the incentive payment (time preference);[70](#_ENREF_70),[71](#_ENREF_71) payments conditional on absolute improvement (on a continuous scale) rather than relative to other peoples’ performance and payments based on measures of process domains rather than outcome.

The variable ‘who receives the incentive’ was not statistically significantly associated with effectiveness. Theory predicts that paying groups could bring about better performance than paying individuals, because organisations are capable of promoting behaviour change in employees through a wide range of strategies.[72](#_ENREF_72) Our findings might reflect that in many P4P schemes where incentives were paid to groups, individual clinicians benefitted as well. This agrees with two narrative reviews, which found that incentives aimed at the individual provider level and/or team level generally reported positive results compared to very large groups.[1](#_ENREF_1),[69](#_ENREF_69)

Our findings are in line with literature suggesting that the effectiveness of P4P might be greater for process indicators compared to outcome indicators,[2](#_ENREF_2),[73](#_ENREF_73) they are easier for providers to influence directly and less likely to be influenced by external factors than are health outcomes. Even in large and relatively homogeneous schemes (in terms of design features), such as the UK quality and outcomes framework (QOF), higher overall performance on QOF indicators is not associated with lower mortality rates for key incentivised conditions.[74](#_ENREF_74)

The effectiveness of interventions evaluated with poorer designs such as inadequate control groups is likely to be over-estimated.[11](#_ENREF_11),[12](#_ENREF_12),[75](#_ENREF_75) Adequate controls are important because incentive schemes often are accompanied by other improvement activities and additional funds. These findings demonstrate the need for more rigorous evaluations, with longer term follow up to assess sustainability of effect.

Other factors, not examined in this paper, will also influence the results of P4P and explain heterogeneity, such as setting, length of programme, organisational preparedness, healthcare system, provider/patient characteristics, dimension of care.[1](#_ENREF_1),[69](#_ENREF_69),[76-79](#_ENREF_76) These require mixed methods evaluative approaches.

**Implications**

Despite the popularity of P4P in healthcare, their effectiveness in improving health care performance has been shown to be variable and modest (particularly in improving health outcomes). Evaluations are often poorly designed and lack adequate controls and so over-estimate their impact. Some of the heterogeneity in results can be explained by design features. Policy makers should design their schemes taking into account how these features enhance effectiveness. These include payment of sufficiently large incentives and the use of performance measurement and payment regimes which reduce the perceived risk of not being paid if the required performance improvement is made. Evaluations of P4P should use well controlled designs and include assessment of their cost-effectiveness.

1. Van Herck P, De Smedt D, Annemans L, Remmen R, Rosenthal M, Sermeus W. Systematic review: Effects, design choices, and context of pay-for-performance in health care. *BMC Health Services Research* 2010; **10**(1): 1-13.

2. Hamilton FL, Greaves F, Majeed A, Millett C. Effectiveness of providing financial incentives to healthcare professionals for smoking cessation activities: systematic review. *Tob Control* 2013; **22**(1): 3-8.

3. Oxman AD, Fretheim A. Can paying for results help to achieve the Millennium Development Goals? A critical review of selected evaluations of results-based financing. *J Evid Based Med* 2009; **2**(3): 184-95.

4. Allen T, Mason T, Whittaker W. Impacts of pay for performance on the quality of primary care. *Risk Management and Healthcare Policy* 2014; **7**: 113-20.

5. Scheffler RM. Pay For Performance (P4P) Programs in Health Services:What is the Evidence? [http://www.who.int/healthsystems/topics/financing/healthreport/P4PWHR2010ShefflerFINAL.pdf:](http://www.who.int/healthsystems/topics/financing/healthreport/P4PWHR2010ShefflerFINAL.pdf%3A) World Health Organisation, 2010.

6. Eijkenaar F. Pay for performance in health care: an international overview of initiatives. *Med Care Res Rev* 2012; **69**(3): 251-76.

7. Shea B, Grimshaw J, Wells G, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology* 2007; **7**(1): 10.

8. Ogundeji YK. Pay for Performance for Health service providers. [http://etheses.whiterose.ac.uk/9036/:](http://etheses.whiterose.ac.uk/9036/%3A) University of York; 2015.

9. Witter S, Fretheim A, Kessy FL, Lindahl AK. Paying for performance to improve the delivery of health interventions in low- and middle-income countries. *Cochrane Database Syst Rev* 2012; **15**(2).

10. Gillam SJ, Siriwardena AN, Steel N. Pay-for-performance in the United Kingdom: impact of the quality and outcomes framework: a systematic review. *Ann Fam Med* 2012; **10**(5): 461-8.

11. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin; 2002.

12. Tilling K, Sterne J, Brookes S, Peters T. Features and designs of randomized controlled trials and non-randomized experimental designs In: Bowling A, Ebrahim S, eds. Handbook of health research methods: Investigation, Measurement and Analysis. Berkshire, England: Open University Press, Mcgraw-Hill; 2005.

13. Bland M. An introduction to medical statistics. 3rd ed. ed. Oxford: Oxford University Press; 2000.

14. Borenstein M, Hedges LV, Julian PTH, Rothstein HR. Introduction to meta-analysis. Chichester, West Sussex, U.K.: John Wiley & Sons; 2009.

15. Moerbeek M. The consequence of ignoring a level of nesting in multilevel analysis. *MULTIVARIATE BEHAVIORAL RESEARCH* 2004; **39**(1): 129-49.

16. Snijders TAB, Bosker RJ. Multilevel analysis : an introduction to basic and advanced multilevel modeling. Los Angeles, London: SAGE; 2012.

17. Luke DA. Multilevel modeling. Thousand Oaks, Calif. London: SAGE; 2004.

18. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med* 2000; **19**(13): 1707-28.

19. Higgins JP, Green S. Cochrane handbook for systematic reviews of interventionsVersion 5.1.0. 2011 (accessed 20 June 2013).

20. Harbord RM, Higgins JPT. Meta-regression in Stata. *Stata Journal* 2008; **8**(4): 493-519.

21. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* 1999; **18**(20): 2693-708.

22. Gelman A. Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *American Statistical Association and the American Society for Quality* 2006; **48**(3): 432-5.

23. Kouides RW, Bennett NM, Lewis B, et al. Performance-based physician reimbursement and influenza immunization rates in the elderly. *American Journal of Preventive Medicine* 1998; **14**(2): 89-95.

24. An LC, Bluhm JH, Foldes SS, et al. A randomized trial of a pay-for-performance program targeting clinician referral to a state tobacco quitline. *Arch Intern Med* 2008; **168**(18): 1993-9.

25. Basinga P, Gertler PJ, Binagwaho A, Soucat ALB, Sturdy J, Vermeersch CMJ. Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. *The Lancet* 2011; **377**(9775): 1421-8.

26. Salize HJ, Merkel S, Reinhard I, Twardella D, Mann K, Brenner H. Cost-effective primary care-based strategies to improve smoking cessation: more value for money. *Arch Intern Med* 2009; **169**(3): 230-5.

27. Twardella D, Brenner H. Effects of practitioner education, practitioner payment and reimbursement of patients' drug costs on smoking cessation in primary care: a cluster randomised trial. *Tob Control* 2007; **16**(1): 15-21.

28. Bardach NS, Wang JJ, De Leon SF, et al. Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: a randomized trial. *Jama* 2013; **310**(10): 1051-9.

29. Glickman SW, Ou FS, DeLong ER, et al. Pay for performance, quality of care, and outcomes in acute myocardial infarction. *Jama* 2007; **297**(21): 2373-80.

30. Cutler TW, Palmieri J, Khalsa M, Stebbins M. Evaluation of the relationship between a chronic disease care management program and california pay-for-performance diabetes care cholesterol measures in one medical group. *J Manag Care Pharm* 2007; **13**(7): 578-88.

31. Rosenthal MB, Frank RG, Li Z, Epstein AM. Early experience with payfor- performance: From concept to practice. *JAMA* 2005; **294**(14): 1788–93.

32. Lee TT, Cheng SH, Chen CC, Lai MS. A pay-for-performance program for diabetes care in Taiwan: a preliminary assessment. *Am J Manag Care* 2010; **16**(1): 65-9.

33. Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M. Effects of Pay for Performance on the Quality of Primary Care in England. *New England Journal of Medicine* 2009; **361**(4): 368-78.

34. Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M. Reduced mortality with hospital pay for performance in England. *The New England journal of medicine* 2012; **367**(19): 1821-8.

35. Jha AK, Joynt KE, Orav EJ, Epstein AM. The Long-Term Effect of Premier Pay for Performance on Patient Outcomes. *New England Journal of Medicine* 2012; **366**(17): 1606-15.

36. Lindenauer PK, Remus D, Roman S, et al. Public reporting and pay for performance in hospital quality improvement. *The New England journal of medicine* 2007; **356**(5): 486-96.

37. Kontopantelis E, Reeves D, Valderas JM, Campbell S, Doran T. Recorded quality of primary care for patients with diabetes in England before and after the introduction of a financial incentive scheme: a longitudinal observational study. *BMJ Quality & Safety* 2012.

38. Cupples ME, Byrne MC, Smith SM. Secondary prevention of cardiovascular disease in different primary healthcare systems with and without pay-forperformance. *Heart* 2008; **94**(1594).

39. Kuo RN, Chung KP, Lai MS. Effect of the pay-for-performance program for breast cancer care in Taiwan. *Am J Manag Care* 2011; **17**(5 Spec No): e203-11.

40. Li YH, Tsai WC, Khan M, et al. The effects of pay-for-performance on tuberculosis treatment in Taiwan. *Health policy and planning* 2010; **25**(4): 334-41.

41. Gilmore AS, Zhao Y, Kang N, et al. Patient outcomes and evidence-based medicine in a preferred provider organization setting: a six-year evaluation of a physician pay-for-performance program. *Health services research* 2007; **42**(6 Pt 1): 2140-59.

42. Chien AT, Li Z, Rosenthal MB. Improving timely childhood immunizations through pay for performance in Medicaid-managed care. *Health services research* 2010; **45**(6 Pt 2): 1934-47.

43. McMenamin SB, Schauffler HH, Shortell SM. Support for smoking cessation interventions in physician organizations: results from a national study. *Med Care* 2003; **41**(1): 396-406.

44. Kruse GR, Chang Y, Kelley JH, Linder JA, Einbinder JS, Rigotti NA. Healthcare system effects of pay-for-performance for smoking status documentation. *Am J Manag Care* 2013; **19**(7): 554-61.

45. Chen JY, Tian H, Taira Juarez D, et al. The effect of a PPO pay-for-performance program on patients with diabetes. *Am J Manag Care* 2010; **16**(1): e11-9.

46. Gavagan TF, Du H, Saver BG, et al. Effect of Financial Incentives on Improvement in Medical Quality Indicators for Primary Care. *J Am Board Fam Med* 2010; **23**: 622– 31.

47. Tsai WC, Kung PT, Khan M, et al. Effects of pay-for-performance system on tuberculosis default cases control and treatment in Taiwan. *J Infect* 2010; **61**(3): 235-43.

48. Doran T, Kontopantelis E, Valderas JM, et al. Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework; 2011.

49. St Jacques PJ, Patel N, Higgins MS. Improving anesthesiologist performance through profiling and incentives. *J Clin Anesth* 2004; **16**(7): 523-8.

50. Bischoff K, Goel A, Hollander H, Ranji SR, Mourad M. The Housestaff Incentive Program: improving the timeliness and quality of discharge summaries by engaging residents in quality improvement. *BMJ Qual Saf* 2013; **22**(9): 768-74.

51. Chang FC, Hu TW, Lin M, Yu PT, Chao KY. Effects of financing smoking cessation outpatient services in Taiwan. *Tob Control* 2008; **17**(3): 183-9.

52. Coleman T, Lewis S, Hubbard R, Smith C. Impact of contractual financial incentives on the ascertainment and management of smoking in primary care. *Addiction* 2007; **102**(5): 803-8.

53. Calvert M, Shankar A, McManus RJ, Lester H, Freemantle N. Effect of the quality and outcomes framework on diabetes care in the United Kingdom: retrospective cohort study. *BMJ* 2009; **338**.

54. Larsen DL, Cannon W, Towner S. Longitudinal assessment of a diabetes care management system in an integrated health network. *J Manag Care Pharm* 2003; **9**(6): 552-8.

55. Simpson CR, Hannaford PC, Ritchie LD, Sheikh A, Williams D. 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* 2011; **61**(588).

56. Simpson CR, Hippisley-Cox J, Sheikh A. Trends in the epidemiology of smoking recorded in UK general practice. *Br J Gen Pract* 2010; **60**: 121-7.

57. Vaghela P, Ashworth M, Schofield P, Gulliford MC. Population intermediate outcomes of diabetes under pay for performance incentives in England from 2004 to 2008. *Diabetes Care* 2008.

58. Srirangalingam U, Sahathevan SK, Lasker SS, Chowdhury TA. Changing pattern of referral to a diabetes clinic following implementation of the new UK GP contract. *British Journal of General Practice* 2006; **56**(529): 624-6.

59. Fagan PJ, Schuster AB, Boyd C, Marsteller JA, Griswold M, Murphy SM. Chronic care improvement in primary care: evaluation of an integrated pay-for-performance and practice-based care coordination program among elderly patients with diabetes. *Health services research* 2010; **45**: 1763-82.

60. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. ed. Hillsdale, N.J.: L. Erlbaum Associates; 1988.

61. Mittlböck M, Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analyses. *Stat Med* 2010; **25**(24): 4321-33.

62. Kontopantelis E, Springate DA, Reeves D. A re-analysis of the Cochrane Library data: the dangers of unobserved heterogeneity in meta-analyses. *PLoS One* 2013; **8**(7).

63. Chen TT, Chung KP, Lin IC, Lai MS. 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 services research* 2011; **46**(1 Pt 1): 47-60.

64. Pope GC. Overview of Pay for Performance Models and Issues. In: Cromwell J, Trisolini MG, Pope GC, Mitchell JB, Greenwald LM, editors. Pay for Performance in Health Care:Methods and Approaches. North Carolina: Research Traiangle Press; 2011. p. 33-60.

65. Desquins B, Holly A, Huguenin J. Physicians’ working practices: target income, altruistic objectives or a maximization problem? Lausanne, Switzerland: Institute of Health Economics and Management (IEMS), University of Lausanne 2009.

66. Evans RG. Supplier-Induced Demand: Some Empirical evidence and Implications. In: Perlman M, ed. The Economics of Health and Medical Care New York: Wiley; 1974.

67. Meacock R, Kristensen SR, Sutton M. THE COST-EFFECTIVENESS OF USING FINANCIAL INCENTIVES TO IMPROVE PROVIDER QUALITY: A FRAMEWORK AND APPLICATION. *Health Economics* 2014; **23**(1): 1-13.

68. Glasziou PP, Buchan H, Del Mar C, et al. When financial incentives do more good than harm: a checklist. *BMJ* 2012; **13**(345).

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

70. Price C. Time, Discounting, and Value. Oxford: Blackwell; 1993.

71. Loewenstein G, Prelec D. Anomalies in Intertemporal Choice: Evidence and an Interpretation. *The Quarterly Journal of Economics* 1992; **107**(2): 573-97.

72. Stewart R. Management of health care. Aldershot, Hants, England: Ashgate/Dartmouth; 1998.

73. Huang J, Yin S, Lin Y, Jiang Q, He Y, Du L. Impact of pay-for-performance on management of diabetes: a systematic review. *Journal of evidence-based medicine* 2013; **6**(3): 173-84.

74. Kontopantelis E, Springate DA, Ashworth M, Webb RT, Buchan IE, Doran T. Investigating the relationship between quality of primary care and premature mortality in England: a spatial whole-population study. *BMJ* 2015; **350**(H904).

75. Ireland M, Paul E, Dujardin B. Can performance-based financing be used to reform health systems in developing countries? *Bull World Health Organ* 2011; **89**(9): 695–8.

76. Stockwell A. Evaluation of Financial Incentives as a Quality Improvement Strategy in the Public Hospital Context: Clinicians Attituides, Design Variables, and Economic Costs: Queensland University of Technology; 2010.

77. Pierce RG, Bozic KJ, Bradford DS. Pay for performance in orthopaedic surgery. *Clin Orthop Relat Res* 2007; **457**: 87-95.

78. Ssengooba F, McPake B, Palmer N. Why performance-based contracting failed in Uganda – An “open-box” evaluation of a complex health system intervention. *Social Science & Medicine* 2012; **75**(2): 377-83.

79. Ogundeji YK, Jackson C, Sheldon TA, Olubajo O, Ihebuzor N. Pay for performance in Nigeria: the influence of context and implementation on results. *Health policy and planning* 2016.

1. Effect estimates in studies lacking control group were mean change before and after the intervention, and change from baseline trends for interrupted time series designs [↑](#footnote-ref-1)