Health Policy 120 (2016) 1141–1150

Contents lists available at ScienceDirect

Health Policy

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

Review

The effectiveness of payment for performance in health care: A meta-analysis and exploration of variation in outcomes

Yewande Kofoworola Ogundeji[a,b,*], John Martin Bland[a], Trevor Andrew Sheldon[c]

[a] Department of Health Sciences, University of York, York, YO10 5DD, UK
[b] Health Strategy and Delivery Foundation (HSDF), 1980 Wikki Spring Street, Maitama, Abuja, Nigeria
[c] Hull York Medical School, University of York, York, YO10 5DD, UK

ARTICLE INFO

Article history:
Received 27 February 2016
Received in revised form 17 August 2016
Accepted 2 September 2016

Keywords:
Pay-for-performance (P4P)
Heterogeneity
Meta-regression
Design
Evaluation

ABSTRACT

Background: Pay for performance (P4P) incentive schemes are increasingly used worldwide 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 standardized effect sizes (standardized mean difference, SMD) and each study was categorized 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 (I^2) was 99.9%. Estimates of effect of P4P schemes were lower in evaluations using randomized controlled trials (SMD = 0.08; 95% CI: 0.01–0.15) compared to no controls (0.15; 95% CI: 0.09–0.21), 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: 0.06–0.31).

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: 1.07–10.64). 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: 0.62–6.56) and if there is a lower perceived risk of not earning the incentive (OR = 2.9; 95% CI: 0.78–10.83). Schemes evaluated using less rigorous designs were 24 times more likely to have positive estimates of effect than those using randomized controlled trials (OR = 24; 95% CI: 6.3–92.8).

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 behavior.

[*] Corresponding author at: Health Strategy and Delivery Foundation (HSDF), 1980 Wikki Spring Street, Maitama, Abuja, Nigeria.
E-mail address: ykogundeji@gmail.com (Y.K. Ogundeji).

http://dx.doi.org/10.1016/j.healthpol.2016.09.002
0168-8510/© 2016 Elsevier Ireland Ltd. All rights reserved.

1. 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, utilization, 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]. 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 [4,5]. Heterogeneous results observed across P4P schemes might be explained by variation in design features, contexts, implementation factors, and evaluation design between the schemes [1,6]. 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 analyze 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.

2. Methods

We conducted a systematic review and meta-analysis.

2.1. 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], which scored 11/11 on the AMSTAR checklist, conducting the search up until April 2016 (Supplementary files S1–S5) [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 organizations involved in implementing and evaluating P4P were also searched e.g., The World Bank; Global Alliance for Vaccines and Immunizations (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.

2.2. 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 incentivized). 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]. We also categorized the design of each evaluation to indicate whether it was a randomized 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].

2.3. Statistical analysis

2.3.1. 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] Therefore, we converted them into two common measures which could be compared across studies and combined.

First we converted them to standardized 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 [13,14]. 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 [15,16]. Selecting a primary outcome measures from the multiple outcomes reported was difficult, as the indicators/measures incentivized 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] (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) [14,17]. 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

[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.