Reviewer's report

Title: Systematic review: outcomes, mechanisms, and context of pay-for-performance in health care

Version: 1 Date: 3 March 2010

Reviewer: Frank Eijkenaar

Reviewer's report:

1. Title and abstract

1.1. Minor essential revision. The title contains the word ‘outcomes’. The word ‘effects’ is more appropriate in the context of pay-for-performance since ‘outcomes’ often refers to a particular type of performance indicator (or measure) used in pay-for-performance programs. As a result, the title does not reflect appropriately the subject of the paper, i.e. a systematic review of studies evaluating the effectiveness of pay-for-performance, and the influence of mechanisms and context on the results.

1.2. Minor essential revision(s). A few comments on the abstract. First, the objectives of the paper can be stated more clearly. The paper summarizes evidence on P4P effects, as well as on the impact of mechanisms and context on these effects. Second, ‘substantial evidence’ (see under ‘results’) is too strongly expressed; there is a lot of evidence, but most of it should be interpreted with great caution. Third, which are the ‘other quality domains’ on which the evidence of the effect of pay-for-performance is scarce? Fourth, the second sentence under ‘results’: this finding is not discussed explicitly in the results section (it is briefly described in the discussion while it should (also) be described in the results section). Finally, this paper does not provide ‘first indications on how findings are likely to relate to P4P mechanisms and context’. It does provide further indications and strongly contributes to the literature and knowledge in this field.

2. Introduction and background

2.1. Major compulsory revision. The main question posed by the authors is not defined sufficiently well (i.e. the objectives should be stated more clearly). Although it is clear that one objective of the paper is to systematically review existing evidence on the effectiveness of pay-for-performance, from the first section (labeled ‘background’ where ‘introduction’ seems more appropriate) it does not become clear how ‘context’ and ‘mechanisms’ are defined and what will be analyzed exactly. Finally, at the bottom of page 4 the authors talk about a review of ‘requisite conditions’. Is this also an objective of the paper? Or do the authors extract requisite conditions themselves from the things they find in the literature?

2.2. Discretionary revision. The used definition of pay-for-performance is not very
clear (page 4, second sentence). In a broad definition, explicit financial incentives are used to improve quality while containing or reducing costs/volume, without jeopardizing equity and fairness. Strictly speaking, good performance does not only consist of good quality as is stated in the paper; efficiency metrics are also relevant. If the authors only focus on quality, they should state that explicitly in the introduction (the authors do state this more or less in box 1).

2.3. Discretionary revision. The course followed around the world regarding the implementation of pay-for-performance is not necessarily ‘uninformed’ (page 4). For example, theory will also play a role in the development of these programs.

2.4. Discretionary revision. It may not be entirely clear what is meant by ‘acute’ hospital care (one of the inclusion criteria). When is hospital care not acute anymore?

2.5. Minor essential revision. What is exactly the function of references 21-23?

3. Methods

3.1 Major compulsory revision. It is not very clear which time frame is used for the literature search and the discussion of studies; which of the studies are actually discussed in the paper and which are listed in eTable 8 (compare what is said in the abstract and at page 4, with page 5)? Which and how many studies are newly identified in this paper and which were already found by the authors of previous reviews? During inspection of the paper this gradually becomes clear but is would be better if the authors just state this more clearly.

3.2. Discretionary revision. Regarding the search string: perhaps more studies would have come up if the words ‘doctor’ and ‘institution’ were also included.

3.3. Discretionary revision. Note that there is overlap with previous reviews (e.g. Christianson et al. 2008, ref. number 7; Mehrotra et al. 2009, ref. number 13). Both studies searched till summer 2007. This is not stated in the text.

3.4. Discretionary revision. The authors state that more than 60 international experts were consulted. If possible, the authors could include an additional table with the names of these experts. The reader may be interested in this information.

3.5. Minor essential revision. eTable 5 lists 61 references. How does this number relate to figure 1 (the flow chart)?

3.6. Discretionary revision. Page 6, 4th sentence: replace ‘reviews’ with e.g. studies or evaluations.

3.7. Minor essential revision. At the bottom of page 6 the authors state “the following data were extracted and summarized in evidence tables”. Not all this information is in the tables while most of it seems relevant.

3.8. Major compulsory revision. From eTable 6 it does not become clear which the excluded 50 references are (after quality appraisal, see flow chart); there are
only 43 overrules (assuming that ‘overrule’ means excluded based on insufficient methodological quality, see also comment 4.3. below).

3.9. Minor essential revision. How do both eTables 7 relate to each other? What are the differences between the studies in both tables? It looks like if the first table contains cost-effectiveness studies and the second table modeling studies (different criteria are applied). My suggestion would be to state this clearly in the table titles.

4. Results: outcome findings

4.1. Minor essential revision. Page 7: the number of studies per country does not add up to 128, but to 127.

4.2. Major compulsory revision. The section on outcome findings is not a ‘comprehensive overview’, as is stated at the bottom of page 4. Such a number of studies requires a more extensive section on the effects of pay-for-performance. Only a subset of the 128 studies is discussed in the text. The discussion on the impact on clinical practice, practice organization, patient satisfaction, and equity could be structured more clearly as well (perhaps under separate subheadings).

4.3. Major compulsory revision. eTables 3, 5, 6, 7 and 8 have no legend. For example, in eTable 6 under ‘sample size’: what does ‘1’ mean and what does ‘-1’?

4.4. Discretionary revision. Page 7: “some studies took place in both settings”. How many exactly?

4.5. Discretionary revision. When referring to numbers (see e.g. page 7), adhere to one approach. Use letters (‘eighteen’) or numbers (‘18’), but not both.

4.6. Minor essential revision. What is meant by “seven study designs were observed based on randomization, comparison group, and time. Nine randomized studies were included” (bottom of page 7)?

4.7. Major compulsory revision. Explain what is done in table 1, how it is constructed, and why you have chosen for this format (try to ask yourself whether this is the most suitable approach). Report the number of studies that are included in table 1, and how many showed negative and positive results (NB. I counted 40 studies in table 1, this seems to contradict with what is reported at the bottom of page 7). In addition, report how many studies assessed the impact on non-incentivized measures. Finally, studies with poorer methodological quality are also discussed in the text, so there is no clear and separate description of the results presented in table 1. This distinction (studies in table vs. studies not in table) should be stated more clearly. I would recommend also reporting always the total number (and references) of studies with positive, absent, conflicting, and negative effects.

4.8. Minor essential revision. Page 8, last sentence: the observation of a “closing
gap for performance differences” is not followed by (a) reference(s).

4.9. Minor essential revision. Page 9: although stated by the authors, reference 68 is not a Spanish study. In fact, this study was conducted in the US, California. Shouldn’t it be reference 109?

4.10. Major compulsory revision. Page 9: The Kahn et al. (2006) and Fleetcroft and Cookson (2006) references (numbers 310 and 311 in the appendix) are not discussed in the section on cost-effectiveness results in the main text. In addition, the An et al. (2008) reference (number 263 in the appendix) is discussed (ref. number 73), but not listed in eTables 7.

5. Results: mechanism findings

5.1. Discretionary revision. The word ‘context’ (first sentence on page 9) seems a little misplaced since context findings are discussed in the next section. (NB. I do like the idea of discussing the findings within a quality improvement cycle).

5.2. Minor essential revision. This section also discusses findings from (at least 17) studies not presented in table 1 (see comment 4.7). Again this distinction could be stated more clearly.

5.3. Minor essential revision. Under ‘quality goals and targets’: what about structure indicators? In addition, it is to be expected that process indicators showed higher improvement rates than outcome indicators because, as opposed to outcome indicators, process indicators can be directly influenced by health care providers. The result on outcome indicators is largely a result of factors that providers cannot influence (e.g. chance, patient adherence to recommended therapies) and patients’ health status. This is why risk-adjustment and other risk-mitigating measures (e.g. carve-outs, ex-post risk-sharing schemes, exception reporting, etc.) are so important, especially in case outcome indicators are used for (external) provider accountability. Further, is it possible to say something about differences between current pay-for-performance programs in the (mean) number and types of performance indicators used, and on how often other dimensions (e.g. clinical effectiveness vs. patient satisfaction, practice organization, use of IT, teamwork, expenditure, utilization) are targeted? A related point is whether the authors can elaborate on the main diseases/conditions (patient groups) targeted by pay-for-performance and on possible explanations for these observations (e.g. no performance indicators available for other diseases, targeted diseases have the most potential for quality improvement/health gain, etc.).

5.4. Minor essential revision. Top of page 10: “Study findings…influences program impact”. Can the authors provide (a) reference(s) for this observation and a few words on the direction of this influence (positive or negative)?

5.5. Minor essential revision. The 3rd sentence on page 10 requires revision.

5.6. Minor essential revision. The four references cited in the 4th sentence on page 10: the reader will be interested to know which studies did and which
studies did not involve stakeholders in target selection.

5.7. Major compulsory revision. Under quality measurement: what percentage of the studies analyzed reports having used some form of risk-adjustment? And how many studies actually assessed whether gaming occurred (if it is minimally assessed, and it looks like it is, one cannot conclude on whether it has been a problem)? Finally, is this all the authors can say about quality measurement?

5.8. Discretionary revision. Under P4P incentives, page 10: “…is clouded by other factors”. Which factors are meant exactly?

5.9. Discretionary revision. At pages 10 and 11, the authors talk about the QOF and CMS/Premier programs. For some readers this may be confusing since these programs have not been introduced earlier in the paper.

5.10. Major compulsory revision. Under P4P incentives, page 10: the authors suggest that a fixed threshold is used to capture best performers whereas a continuous scale is used to capture best improvers. The word ‘capture’ is a little misplaced here and both methods should not be posed against each other as being two completely different approaches. Both essentially reward absolute performance as opposed to the tournament approach in which relative performance is rewarded. The main difference between both methods is that a fixed threshold rewards achievement while a continuous scale also rewards improvement. A continuous scale can therefore also be used to ‘capture’ best performers. (NB. which of the 4 accompanying references showed positive effects, and which of these no effects?)

5.11. Major compulsory revision. Under P4P incentives: again, it would be interesting to see some figures on the prevalence of particular approaches used. For example, how often does one observe the competitive approach (‘tournaments’) in the literature as compared to non-competitive approaches? What is the mean duration, frequency and size of the incentives and how often are financial penalties part of pay-for-performance programs? What is the most common targeted entity (individual physicians, groups, institutions)? If the authors purposely did not discuss all of these elements for whatever reason, they should state this beforehand.

5.12. Minor essential revision. Middle of page 11: what is meant by “nationwide corrections were required”? And could the authors say anything about a difference in effect of rapidly implemented programs vs. gradually implemented programs?

5.13. Minor essential revision. Are there differences in effects on performance between voluntary and mandatory programs? And is there information on the impact of the finding found in reference 85 (cited at page 11)?

5.14. Minor essential revision. Page 12, 3rd sentence: “…remain mixed”. Are there references for this finding?

5.15. Major compulsory revision. Middle of page 12: reference 109 is a
time-series study, as noticed by the authors. However, this reference is not listed in table 1, despite the fact that the authors stated on page 8 that table 1 also contains identified time-series studies.

5.16. Major compulsory revision. Middle of page 12: the conclusion that “this arrangement seems to have reinforced the P4P effect” is incorrect because one cannot be sure about the contribution of P4P to the observed effect, unless a randomized design was employed with multiple intervention groups and a control group. If not, P4P may not have contributed at all to the observed effect. In practice it is very hard to disentangle the effects of different quality improvement interventions simultaneously in place.

5.17. Minor essential revision. Under evaluation: isn’t there more to say about evaluation? For example, how often do program administrators employ a formal evaluation procedure on a regular basis? Are physicians’ perceptions of the functioning of the program assessed and how are valuable lessons learned adopted in practice? Etc.

6. Results: context findings


6.2. Minor essential revision. Can you give an example of ‘health system values’ and the presence of ‘congruence with health system values’? It is not entirely clear what is meant here.

6.3. Minor essential revision. Bottom of page 12: “…diverse initiatives….” I would replace the word ‘diverse’ by e.g. ‘more fragmented’ or ‘more decentralized’.

6.4. Minor essential revision. First sentence page 13: “This affected…..” What is meant by ‘this’? When you put it like this, the sentence can be interpreted like P4P results affected many of the described mechanisms. However, the authors probably mean that the level of implementation/decision-making affected the mechanisms in place which in turn affected results.

6.5. Major compulsory revision. Regarding healthcare system characteristics: a lot more can be said about this. What about the influence of how healthcare is financed (Bismarck vs. Beveridge), how healthcare is purchased and by whom, is the system market-oriented or highly regulated, etc. These factors may well have (had) an important impact on P4P mechanisms and results.

6.6. Minor essential revision. Last sentence under ‘payer characteristics’: shouldn’t ‘mixed’ be replaced with e.g. ‘largely absent’? The word ‘mixed’ suggests that evidence is present which, according to the authors, is not.

6.7. Major compulsory revision. Is the payment system a characteristic of the payer or the provider? Further, capitation does not only provide incentives for underuse, but also for quality skimping (which may also occur in FFS).
Pay-for-performance is certainly not only used to correct over– or underuse (which seems to be assumed implicitly here), but also simply to improve the quality of the care delivered. Finally, I think the authors may be able to say something about other payer characteristics, e.g. their role in care purchasing and in the design and implementation of P4P programs, the description of the benefit package (which is also a healthcare system characteristic), the extent to which payers bear financial risk (which has consequences for incentives for risk selection), etc.

6.8. Discretionary revision. Under provider characteristics: isn’t a ‘patient-centered culture a part of the organizational culture. Further, what is meant by ‘in terms of participation’ (last sentence of page 13)? Is provider participation in P4P actually part of P4P performance?

6.9. Minor essential revision. With regard to the target unit of the incentive: the reader would be interested to know how these different units are defined. For example, aren’t physician groups or IPA’s organizations themselves? When you consider only hospitals as organizations, the fact that programs directed at organizations often showed no effect may also be a result of such programs being less common than programs directed at physician(s) (groups).

6.10. Minor essential revision. Page 14, 4th sentence: “… additional efforts seemed to be required”. Such as? And are there examples of programs (i.e. references) showing positive results at this level?

6.11. Minor essential revision. Top of page 15: only one study is cited (number 110), so it is not correct to talk about ‘other studies’ (plural).

6.12. Discretionary revision. Regarding the study of Mehrotra, the authors could report the actual median number (which was 39 physicians). And regarding the study of Tahrani, the authors could state whether small or large practices were performing better at baseline.

6.13. Discretionary revision. Under provider characteristics, if data permits the authors could consider to also talk about the influence of characteristics of the provider market, e.g. its competitiveness.

6.14. Major compulsory revision. Under patient characteristics: what about patients’ health status? Patients are not randomly distributed across providers, which makes risk-adjustment a crucial part of each pay-for-performance program. Is there evidence on the effects on (the prevalence of) patient selection based on (indicators of) health status of pay-for-performance programs that do and do not employ risk-adjustment?

6.15. Minor essential revision. The last sentence on page 15 needs revision.

7. Discussion and conclusions

7.1. Minor essential revision. 3rd sentence, page 16: the mere identification of 128 relevant studies is not an illustration of the evolution of the evaluative
pay-for-performance literature-base. Further, the authors speak about the “difference in retrieval number”. How many studies did the authors identify that authors of previous reviews did not?

7.2. Discretionary revision. Middle of page 16: the sentence “the fact that P4P should learn from multiple study designs” requires revision.

7.3. Discretionary revision. On page 16, the authors state that “the scientific quality of the current evidence is fair”. I wonder if other authors (besides those cited) agree with this. “Fair” may still be too strong a word. The vast majority of identified studies was not randomized (only 9 were) and roughly 75 studies were either cross-sectional or employed a simple before-and-after design. I do acknowledge the fact that as the evidence-base continues to grow, conclusions on the effects of pay-for-performance can increasingly be drawn with more certainty, despite the fact that the scientific quality of current evidence is still poor (128 is already quite a number).

7.4. Minor essential revision. Bottom of page 16: the ‘evolution’ the authors talk about is indeed occurring. However, it is not backed by the data presented here.

7.5. Minor essential revision. Page 17, sentences 4-6: these findings are described here explicitly for the first time in the paper (on page 10 a few words are afforded to this implicitly). These findings should, in my view, (also) be addressed explicitly in the ‘results’ section. In addition, perhaps the authors can provide some references after sentence 6.

7.6. Discretionary revision. What is the reason why Petersen et al. (2006) found negative effects, while the authors encounter these rarely?

7.7. Minor essential revision. In the middle of page 17, the authors conclude that gaming and neglecting effects on non-incentivized measures only occurs at a minimal level. It's too early to draw these kind of conclusions because they are based on (discussion of) only a few studies and assessment of a small subset of non-incentivized measures. Furthermore, it is not clear how gaming was assessed in these studies. Based on the evidence, however, there may be some indications of the absence of gaming and a limited neglecting effect on non-incentivized measures.

7.8. Minor essential revision. The first sentence of page 18 needs revision.

7.9. Minor essential revision. What is meant by “follow function” in the Custers et al. citation?

7.10. Major compulsory revision. Some comments regarding the six recommendations: first, explain the rationale behind number 2 and 5. For example, regarding 2, Conrad and Perry (2009) do suggest that adding an outcome indicator may increase the impact of the incentive, but they also stress that some important preconditions must be fulfilled if outcome indicators are used in pay-for-performance programs, including adequate risk-adjustment (see also comment 5.3.). Second, regarding 4, in competitive markets such an approach
may raise anti-trust issues (see the IHA example in California). A uniform set of measures and aggregating patients across payers/purchasers may be allowed, but the design of incentive schemes (including the calculation of bonus payments) can probably not be standardized.

7.11. Discretionary revision. Theory-based recommendation 1: I would add “but keep monitoring scores on old targets to see if achieved results are preserved”.

7.12. Discretionary revision. Theory-based recommendation 2: theory predicts not only that a sufficient incentive size (not too little, but also not too much because this may crowd out physicians’ intrinsic motivation) is needed to support participation, but also to achieve desired results. Compare for example the results of the QOF program in the UK (20% of income) with those of the IHA program in California (2-3% of income). (NB. the last sentence of this recommendation needs revision)

7.13. Minor essential revision. Consider moving the 3-sentence subsection following theory-based recommendation 3 to the top of page 18 (before discussing the 6 evidence-based recommendations). Further, the last sentence of this subsection states that this paper “has completed the contextual framework”. The paper may have added to this framework, but it’s far from completed (see also comments 6.5, 6.7, 6.13, and 6.14).

7.14. Major compulsory revision. Regarding the study limitations: first, what is meant by “restrictions in the databases searched”? Second, what is meant by the first sentence on page 20? Third, regarding the third limitation, see also comment 5.16. Fourth, regarding the data analyzed, there are at least two more limitations: 1) publication bias (especially important in review articles. For example, sponsors of P4P programs yielding negative effects may have objected to publishing the results which may have resulted in an unrepresentative evidence-base of P4P effectiveness), and 2) data quality bias (which makes comparison of results across P4P programs problematic, see also Terris and Litaker 2008). Finally, a limitation of the paper is that potential settings for P4P application were excluded from the study (behavioral health care and nursing home care). Although understandable, it is nonetheless a limitation.

7.15. Major compulsory revision. Finally, in the conclusions section, I miss a conclusion about the overall impact of pay-for-performance on the quality and equity of care, and about the influence of context and mechanisms.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:
I declare that I have no competing interests.