Author's response to reviews

Title: Systematic review: Effects, development choices, and context of pay-for-performance in health care

Authors:

Pieter Van Herck (pieter.vanherck@med.kuleuven.be)
Delphine De Smedt (delphine.desmedt@ugent.be)
Lieven Annemans (lieven.annemans@ugent.be)
Roy Remmen (roy.remmen@ua.ac.be)
Meredith B Rosenthal (mrosenth@hsph.harvard.edu)
Walter Sermeus (walter.sermeus@med.kuleuven.be)

Version: 2 Date: 12 May 2010

Author's response to reviews: see over
Cover letter

Dear Editor, dear reviewers

Thank you very much for providing us with your constructive comments and suggestions. We think the paper has improved a lot due to your input. Detailed feedback also allowed us to improve on logic, structure and writing style.

The most important changes we made are:

- A clarification of objectives
- A clarification/replacement of terms (mechanism, outcome, etc.)
- A more specific, cautious and nuanced approach dealing with the strength of findings
- Correction of some inconsistencies in references and numbers

Below you find a detailed overview of all changes made. Is it possible to share all of these comments and our replies as a whole with the three reviewers? We sometimes refer in our reply to similar suggestions made by the other reviewers.

Two additional files have been added to the supplementary material: one excel formatted large evidence table (a descriptive data extraction overview, not our more detailed analyses) and one report which also provides additional details on the included P4P programs (Annemans et al, 2009).

The reason for these additions concerns multiple requests from the reviewers to provide additional details and to go further into depth with regard to analyses of dominant payment system effects, targeting of other dimensions, the incentive size, etc. To us these requests indicate that our review covers many interesting aspects needing further in depth investigation and summarization. It is a positive sign about the rich data we have gathered. We have added additional information wherever possible within the scope of the paper. However, addressing all requests is not feasible within the limitations of writing space of one paper. Some of those very useful suggestions deserve a separate treatment in more focused papers, for example applying meta analysis to disentangle some conflicting evidence further. We think this goes beyond the scope of the current paper, which is already very broad. In the overview below these issues have been highlighted in bold. Do not hesitate to let us know how we should proceed further on these points (or on any other part of the paper).

Thank you again for all of your useful contributions.
Detailed response to reviewers’ comments and suggestions

Reviewer Frank Eijkenaar

1. Title and abstract
1.1. The authors agree. The term ‘outcome’ has been replaced with ‘effects’, both in the title and throughout the text (except where it is mentioned as type of indicators/measures).
1.2. Suggestions incorporated (see exception 1.2.4 below):
1.2.1. The objective is reformulated as suggested
1.2.2. The term ‘substantial’ is replaced with ‘a large body of’ and ‘to be interpreted with caution’.
1.2.3. ‘other quality domains’ is further specified.
1.2.4. Since this sentence expresses a subjective appreciation the authors feel it would not fit within the main results section of the paper (in addition to the discussion section where it is mentioned). The abstract section ‘results’ implicitly also includes some discussion, since there is no separate heading for the latter provided in BMC author instructions (within conclusions this sentence would be positioned too strongly). However, to make the distinction with results more clear ‘were’ is replaced with ‘can be judged’.
1.2.5. ‘First’ is replaced with ‘further’ as suggested.

2. Introduction and background
2.1. The heading ‘background’ has been chosen to comply with author instructions. ‘Introduction’ would be equally appropriate. This is a decision for the editors to make (left unchanged).
2.1.1. Three objectives have been stated more explicitly, building further on the phrasing suggested above. The term ‘mechanism’ has been replaced with ‘Development choices’ and is immediately further situated. The term ‘context’ is also presented more concretely in terms of subgroups of characteristics. (see changes in last paragraph of background section)
2.1.2. The ‘requisite conditions’ refer to the results addressing the second and third objective. We hope this has been clarified by an immediate explicit statement of those objectives after this sentence.
2.2. Discretionary revision has been applied. The suggested clarification of P4P as concept has been added.
2.3. Discretionary revision. Left unchanged. We think the word ‘uninformed’ right after the sentence about lack of evidence, mixed evidence… is clear when read in this context. We agree theory will also play a role (this topic is covered in another upcoming paper. It reports (roughly speaking) that with regard to theory, programs do not much better than with regard to evidence incorporation). We think this is beyond the scope of this paper which focuses on evidence.

2.4. The word ‘acute’ has been removed before ‘hospital care’, as suggested.

2.5. References 21-23 have been omitted. They indeed served little purpose.

3. Methods

3.1.1. We clarified the time frame used for literature search and studies discussed: (1) no time restriction on inclusion of all relevant studies discussed in other reviews, (2) 2004-2009 July time restriction for electronic databases, (3) no time restriction for reference screening, forward citation and expert input. Using this combination we have an overview of the period 1990-2009, but we acknowledge that for the first period, before 2004, not all search methods have equally been applied (looking at the starting numbers of database input, I’m sure you would agree that such an approach is not feasible). Added ‘without any time restriction’ in the text to a number of the methods so that this would be more clear.

3.1.2. eTable 8 presents all 128 studies. These are all used as input for results on the second and third objective (wherever relevant findings are reported). For the first objective (effects) a selection based on study design was made (Table 1). Added ‘47 studies in total’ to the effect results section to make the distinction more clear.

3.1.3. Added in Description of studies (first result section): 128….’of which 79 were not covered in previous review papers’. A total of 49 studies are covered previously.

3.2. Discretionary revision. The terms ‘doctor’ and ‘institution’ could indeed have been added to our search string. However, this would have reduced our starting number of hits instead of increasing it in the majority of databases. Adding of terms makes the output more specific (things without doctor or institution might have been skipped). We chose for an approach as comprehensively as possible. However, for the databases Econlit and Psychinfo, where we used an additional healthcare setting limitation, there your suggestion could indeed have increased the retrieval number. Because our search strategy is already very extensive, we chose not to add those additional terms.
3.3. You are right in noticing an overlap. The reason is that our search strategy (commencing from 2004 as retrieval year in databases) was already started before the two latest systematic review papers were published. We added them, but it would have been a waste to throw away the (overlapping) work of searching we already performed. Looking and comparing final retrieval numbers between their and our paper, it seems to be the right decision. We were able to fill up some of their gaps. We, as authors, feel that these practical details are a bit beyond what is interesting for the reader to be informed about.

3.4. We asked all experts whether we could publish such a list of their names. However, some of them declined. This might seem a bit strange, but it apparently has to do with some experts working for state institutions (health ministries), having strict regulations on confidentiality, under control of the state employer as owner of intellectual property. They explicitly stated that they wanted to assist in publication retrieval, but their names should not be mentioned. Hence our decision not to include such a list.

3.5. There is indeed a discrepancy between eTable 5 and the flow chart. The difference is explained by papers that were conceptual in nature. Initially we retained a larger group of papers, because we also kept conceptual and P4P design papers without evaluation. Later on we decided to delimit our focus further, since we found so many empirical papers. This means that the flow chart is completely correct (those papers are excluded based on the ‘design’ criterion), but as a footnote of eTable 5 we added that in addition to the table 151 papers are further excluded based on design. We do not list those papers in detail, because they are out of the scope of the paper, focusing on evidence.

3.6. As suggested, ‘reviews’ has been replaced with ‘review papers’. We do not use ‘studies’ or ‘evaluations’, because this might be confusing when dealing with the primary studies afterwards.

3.7. As another potential supplement we have added our complete evidence table, which is an excel sheet for you, reviewers to have a look at. This might be valuable additional information. It is however, dealing with the size and details, impossible to refit all these data into the existing supplement in word format. We are happy to share this additional supplement in excel format. This is a decision for the editors to make.

3.8. Our apologies for this omission. This is clearly an important explanation/legend that we failed to include. Next to overrules we applied a rule of having a minimal score of 8/14. In eTable 6 you can check that next to 43 overrules, 6 papers have below 8 scores. In addition one study is reported in two publications (fourth row of eTable 6). It has been
dealt with as one hit when constructing the flow chart (one additional exclusion). We have added a legend to eTable 6 (see your comment below).

3.9. Suggestion followed: the distinction has been made in the titles.

4. Results: Effect findings

4.1. We forgot to enter one study (Kahn et al, 2006) into eTable 8, and into the description part. This is corrected, also in the design frequency, etc.

4.2. Agreed. We deleted ‘comprehensive’ before ‘overview’. The subset discussed is clarified in number (‘47 studies in total’, see above). Separate subheadings have been added.

4.3. Legends have been added to all these tables, except to eTable 5 (exclusion criteria for relevance screening). This table is self explanatory, no symbols used.

4.4. Specified: Thirteen studies.

4.5. The following rule has been applied: Numbers below ten are written out (letters), numbers above ten are presented as figures. Numbers at the beginning of a sentence are also written out. These rules were applied by an external editing firm (Exact Science).

4.6. Attempt to clarify this by using ‘can be distinguished’ instead of ‘were observed’. ‘criteria’ has been added to clarify the role of randomization, comparison group and time in this sentence. A ‘::’ at the end makes it more clear that the sentence introduces the list that follows.

4.7. Conforming with author instructions, the legend of Table 1 was added at the end of the main manuscript. There you can find the explanation of abbreviations, etc. As you suggested, we added an explanation of the Table in the main text. There we also added the numbers you suggested (see below for positive vs. negative findings exception, 4.7.1.3.). In the Table itself Preventive care was more clearly distinguished.

4.7.1. Issue of number of studies in the clinical effectiveness table. The results of Ritchie et al (1992) (time series) were not included in the Table, because immunization rate was expressed in terms of number of practices and not in terms of patient rate. The study does provide input for two other main sections (objective 2 and 3) and was therefore generally still included. Similarly Langham et al (1995) (concurrent + historic) only reports effects verbally, but provides input for equity. O’Malley et al (2007) (time series) provides association measures, no effect sizes; is also providing input for objective 3. Ashworth et al (2008) (time series) report their findings focusing on equity. MacBride-Stewart et al (2008) (time series) focuses on associations. Millett et al (2008a) (time series) focuses on equity. Tahrani et al (2008) (time series) only
reports an association. We cannot exclude these studies, since they are relevant elsewhere. We think the detailed reasons are not interesting for the reader to know, but we understand that the selection procedure for table 1 should be more explicitly stated. ‘Reporting a clinical effect size’ has been added when introducing the table.

4.7.1.2. Reference 36, Sutton et al (2009) is not included in the Table, because its finding crosses a number of combined patient groups, but not the general population. This makes it practically difficult to include. Therefore we chose to describe this result after the table. We added the combination of patient groups discussed in this study in the text.

4.7.1.3. We acknowledge that Table 1 is a complex representation, since you have to make the combination patient group, study, specific target. However, this representation to our opinion fits best with the real life effects of P4P which are complex and diverse. I understand it is desirable to know how many positive and negative findings there are, going through this table and mentioning this up front. We avoid doing so, because it might be misleading. For example, most of the studies report positive findings on target A, an absent finding on target B, and (very rarely) a negative one on C (There is variability across and within most studies). This is understandable, since a study looks sometimes at 146 measures (UK example). The purpose of this part is partially also a confrontation for the reader with this complexity. Another option would be to present the number of positive and negative findings per target instead of study, but then we would have to deal with the issue of similar measures across different patient groups. This impediment also fits with reality. An attempt to provide numbers by patient group, study and/or target in terms of positive, negative... within the text led to writing out the table almost completely. We do not have the space to do so, nor does it offer more information than presented in the table. However, further suggestions are certainly welcome on this point!

4.7.1.4. The number of studies assessing the impact on non-incentivized measures is added, as suggested.

4.7.1.5. Issue of distinction Table/text + poorer methodological quality also discussed in this section? With regard to clinical effectiveness the text only discusses studies included in the table (exception Sutton et al, see above 4.7.1.2.). For those studies the methodological quality is clear (three study designs selected + having passed quality appraisal). Concerning coordination and continuity of care, and patient centeredness results the study designs are always mentioned when describing results in the text.
With regard to equity of care both the number and the types of research designs are added in the text. In addition a warning sentence was added saying that equity was never investigated as part of randomized studies. Concerning modelling/cost effectiveness effects we added a statement that the few studies vary in methodological quality, but show similar positive results.

4.8. The sentence on the closing gap is based on the 28 studies referenced immediately before it. These studies show a constant and/or decreasing difference between age... groups. The exception with regard to gender is mentioned in the text afterwards. Added: ‘throughout these studies’ to make the link with the previous sentence.

4.9. Corrected in the text: one Spanish study and one study in the US (Rodriguez et al).

4.10. Added one sentence on the finding from Kahn et al (2006). Fleetcroft et al (2006) has been added to the health gain reporting sentence. An et al (2008) has been left unchanged in the cost effectiveness results. They report marginal cost/unit of gain results and compare these with other types of interventions. However, this is a smaller part of a larger randomized study. Hence its main inclusion in eTable 8 as a randomized study instead of primarily defining it as a cost effectiveness study.

5. Results: Development choices (instead of mechanism findings)

5.1. Changed as suggested.

5.2. Added one sentence to clarify the broader inclusion of studies as an input.

5.3. Structural indicators are presented in a minority of studies, which were chosen not to include in this review (other than the 128 studies). The reason is they deal with aspects such as IT improvement, the use of reminder systems, etc. Although interesting, lack of space obliged us to remove these results. A second reason was their presence in only weak study designs. Their removal is a practical choice, after the length of the paper was widely expanding.

5.3.1. We agree on your process vs. outcome analysis, and on the importance of risk adjustment. In the results section we just state the difference in relative improvement as a fact. Space limitations have kept this discussion point later on minimal. I suggest we might add it to the discussion, if after complete revision there is still some room left. Let us know whether this according to your opinion is really essential to discuss or not. To us it seemed obvious, but maybe it is not. The suggestion on number of indicators has been briefly mentioned as part of the discussion section, reporting the general trend.
5.3.2. These suggestions all seem worthwhile, but again, we already present a lot of information in the paper as it is. We have to be careful not wanting to cover everything, P4P in all its details and dimensions, in one paper. The topic is too complex to achieve this. Some of these aspects may however be interesting to cover in a second, more focused paper afterwards. That would provide the space needed to go that far. **Can you agree with us that we selected the main priorities to present? Which of these (and some of the following) points are really crucial to add?** Your suggestion on patient group coverage has been reported implicitly as part of the previous section (see clinical effectiveness and related table, focus on preventive – acute – chronic care, etc.).

5.4. Four references added and direction specified.

5.5. Sentence modified (‘in this area’, ‘such’).

5.6. The four references have been more clearly positioned with regard to level of involvement.

5.7. The main issue we had with risk adjustment identification was although in most studies it is clear that risk adjustment has been used for study purposes, it almost always remains unclear whether that such risk adjustment was part of the P4P measurement scheme itself or not. So, we hesitate a bit to put a number on it, because this might be misleading. We first had a note in our discussion on this point, but it must have been skipped to reduce the length of the paper. **If you think this remark is crucial, we will reintegrate it** (Again, let’s prioritize, because limitations on length of the paper should also be respected).

5.7.1. Added ‘although only three studies measured gaming specifically’ to the point about gaming. We say it’s minimal in these studies, we’re not saying that in general it is no problem. One could again start a whole debate about the gaming comparison point throughout different health systems. How much gaming occurs in non P4P payment systems? Not introduced because of space limitations.

5.7.2. What other key issues with regard to quality measurement would you prioritize to discuss? (We could write a book about it, but those points seemed less important information than the ones we included)

5.8. Added: clouded by e.g. incentive size, level of stakeholder involvement

5.9. QOF and Premier have been reformulated in more general terms to avoid confusion.

5.10. Replaced the ‘capture…’ by ‘quality target achievement and/or improvement’. This seems to fit better with your explanation. We clarified the reference positioning.
5.11. A great deal of this additional information is available through KCE report 118A Annemans et al, 2009 (online available in English at the KCE website). Would it be alright to refer to such a report for additional information? If the KCE and the journal agree, an additional copy could be put in the online supplementary material. We’ve added a copy in the upload for you to have a look at. See Chapter 4.

5.12. Added: ‘in terms of indicator selection, threshold definition and the bonus size per target’. Your second question is difficult to answer. It depends on how you define a gradual implementation, based on:

- Use of pilot studies, demonstration projects or not (positive impression, but clouded)
- Starting with a low number of indicators and then increasing (negative)
- Starting with a low incentive amount and then increasing (negative)
- Using pay for reporting before pay for performance (not clear)
- Etc.

As you can see, this is again a discussion in its own. Is this a key priority to answer our research objectives? Or better to refer to Annemans et al (2009) to guard the length of the paper?

5.13. As stated in the text, the evidence on voluntary vs. mandatory is conflicting. It is our own impression that voluntary leads to better results than mandatory, but this is again clouded by the effect of incentive size, etc. At present we cannot make a harder statement about this issue.

5.14. Two references added.

5.15. Pedros et al (2009) report an improvement in the degree of reporting by providers of adverse drug reactions. This is no aspect of clinical effectiveness and was therefore not included in Table 1. It could be described under a separate Patient Safety heading. However, a number of experts stated problems with the adverse drug reaction outcome instead of a more regular adverse event outcome. We avoided the discussion by not including a separate Patient safety subheading. The main reason is that with the exception of this one study, not much of safety results are available. As you can see in the study objectives, patient safety is not included in the areas focused upon.

5.16. We agree. Your reasoning has been added to the text itself as a word of caution.

5.17. Physician perceptions are treated in another upcoming paper. How often evaluation is applied is again not clear, because papers rarely make a distinction between the
evaluation as part of the study and evaluation as part of the program itself (see also the similar point above on risk adjustment).

6. Results: context findings

6.1. Modified within the text.

6.2. We deleted the part on congruence with health system values, because it was too difficult for us to substantiate. It was too much of an impression without true evidence.

6.3. Modified as suggested

6.4. Clarified ‘this’: ‘The level of decision making…’

6.5. Probably all those health system characteristics will influence P4P development and effects. But can we say that just because right now the UK, Spanish and Australian results are generally (with some mixes on certain targets) a bit more clearcut positive than results from the US, this would be related to the system of social security, type of purchasing or degree of regulation/free market? This would be dangerous to state, especially since so many other variables mediate results and are likely to have a larger impact. Things like the dilution effect are clearer and are reported in the paper. For the other aspects we just don’t have a sufficient number of different healthcare systems in our sample of studies to make further distinctions. However, if crucial, we could add this as a limitation in the discussion section.

6.6. Modified as suggested.

6.7. You can place it under both (in the first version it was under healthcare system characteristics). The important thing is that it is discussed somewhere. This is the case.

6.7.1. **We are not sure what you mean in your second sentence about under- and overuse versus quality (?).** Within quality you have different domains (e.g. IOM). The minimal domain on which P4P focuses is clinical effectiveness. This is about improving health gain and/or following best practice (general clinical quality definition). Evidence based corrections of underuse and/or overuse are the manifestation of the latter, and may lead to the former. I don’t see how you would increase health gain by means of clinical effectiveness in another way through P4P. **Can you explain your reasoning a bit further? You mean by its influence on other domains, on quality skimping…? Then quality skimping would lead to inequity of care, no (which has not been reported)?**

6.7.2. The studies we included do not report useful information on the payer’s role in care purchasing, the benefit package, the extent of financial risk. These aspects would be
interesting, but they are presently not cover in our sample of empirical studies. If we overlook key findings in this regard, please let us know.

6.8. We used the term ‘patient centered culture’ to stay as closely as possible to the original study. It would be risky to change this phrasing, because we might influence the original meaning/definition of concept. Corrected: Deleted ‘in terms of participation’

6.9. We clarified the levels definition by more explicitly mentioning practices and hospitals. The difference in performance is more clear on those levels as compared to others. For IPA’s and medical groups this is less true (although theoretically you would expect a similar kind of distance between incentive receiver and behavioural agent as in the case of a hospital).

6.10. Added: in terms of support generation and/or internal incentive transfer. Clarified in the text that the following sentence is about the same set of references.

6.11. Changed into ‘one study’

6.12. Added: ‘which was 39’. Added a clarification of the historical difference between small versus large with regard to performance.

6.13. As was the case for payer characteristics, as far as we are aware competitiveness is not addressed in P4P evaluation studies.

6.14. In our previous version we briefly referred to the equity of care findings in the previous section, because differences in health status are addressed there (e.g. number of comorbidities). Should we reintegrate such a reference toward this section?

6.14.1. The only evidence available on this point has been provided by Doran et al in a few studies. However, they specifically focus on exception reporting as one form of patient selection prevention. Gravelle et al report similar findings. Right now these aspects are treated very shortly under ‘gaming’ within ‘quality measurement’. We could expand a bit further on this point. Do you suggest placing it here or under the gaming topic?

6.15. Corrected.

7. Discussion and conclusions

7.1. Changed into: ‘the publication rate as described for the last 20 years’. This refers to the description of studies section with a year by year comparison of number of publications. Added: 79 studies were not reviewed previously.

7.2. The sentence has been modified: ‘The use of multiple study designs to….’

7.3. ‘Fair’ has been changed into ‘increasing’. An additional sentence was added to express the nuanced view you expressed, and which we share.

7.4. Sentence rephrased accordingly.
7.5. Do you refer to the following paragraph: ‘Interpretation of effect size is dependent on the primary mission of P4P. When it functions to support uniform minimal standards, P4P serves its purpose in the majority of studies. If P4P is intended to boost performance of all providers, its capability to do so is confirmed for only a number of specific targets, e.g., in diabetic care.’? (as you also mentioned about the abstract)

7.5.1. On this point we currently disagree with you, because this is clearly a subjective interpretation by us, as authors, of the data/results. The sentence itself is about how you look at, how you value data. This kind of subjective statement, according to our opinion, is specifically reserved for the discussion section. Why would you place it in the results section? It’s an appreciation of the complete pattern of clinical effectiveness findings. So, it’s also difficult to provide some references.

7.6. Petersen et al (2006) also encountered negative results only rarely. They did focus more on it, as a warning, because they were the first ones to report negative findings in a review paper. Here we only mentioned in the text that we find the same thing.

7.7. Added: ‘although it is only addressed in a minority of studies’, ‘multiple’ replaced by ‘a few’. We also added your warning sentence about too early conclusions in the text itself.

7.8. Modified

7.9. Changed into: ‘dependent on its objectives and contextual characteristics’

7.10. Added rationale for 5th recommendation: ‘The evidence shows that both may be effective when developed appropriately. A combination of both is most likely to support acceptance and to direct the incentive to both low and high performing providers.’

7.10.1. Added your remark about preconditions to the 2th recommendation

7.10.2. 4th recommendation: We are aware of the antitrust issues which might be raised by a uniform approach across payers/purchasers. This seems to be, for the most part, a typically US situated problem, where competitiveness stands central. It is not an issue in the majority of other Western health systems. Without starting a debate of which balance fits best (probably context specific), we do state that in long term health systems can change. Current health reform in the US is also an example of a fundamental shift. This already brings the US health system closer to many other health systems. Added: ‘although in short term such an approach might raise anti-trust issues in market driven health systems, in long term a rebalancing may become possible.’

7.11. Added as suggested


7.13. Moved as suggested. Changed into ‘further contributed to’ instead of ‘completed’.
7.14. We clarified the first limitation in the text (number of databases consulted).
7.14.1. Sentence about the RCT only approach clarified in the second limitation.
7.14.2. We added a limitation concerning publication bias and data quality bias. Suggested reference added.
7.14.3. We added the limitation with regard to the settings included.
7.15. Added a sentence about clinical effectiveness effects to the conclusion + general statement about quality, the influence of development choices and context.

Reviewer Sukyung Chung

1. The point currently discussed in the section about health system characteristics focuses on avoiding dilution effects by fragmented decision making across P4P programs. We do not agree that this is uninformative. Especially for the US this is of specific relevance, although maybe difficult to change, as one reviewer above noted. Therefore, we did not replace this first finding by a more central focus on dominant payment system effects.
   a. We do agree that dominant payment system effects might be highly influential. It is true that right now the review reports a very cautious statement concerning this point, saying findings are mixed. Based on your suggestion we made a preliminary analysis of what might be possible to insert, to fill this gap. However, only for the UK and Dutch studies as a whole the dominant payment system is clear. For Australia, Spain, Italy, and Germany nothing is mentioned in the studies themselves. Concerning US studies (n=63) only eight studies mention capitation as the dominant payment system and three studies mention FFS. So, only for the US there is a gap to fill for 52 studies. About other types of payment constellation, beyond capitation vs. FFS, even less is mentioned (one study in a salary context, one referring to PPS with DRG use/case rates in a hospital setting). Especially because in the US payment is such a heterogeneous constellation we as authors think it would be unwise to assume certain dominant payment characteristics for specific studies, based on the study being conducted for example within the setting of a hospital vs. in primary care. We therefore would be obliged to gather additional information from the authors of primary studies. In addition, some of the other findings of our review do appear to relate to country borders. Typical examples are the use of very small incentive sizes, the occurrence of dilution, the use of a tournament approach, the use of closed budgets from the start, the
greater attention for cost and productivity, etc. Currently our analysis shows that much of these mediators, for which the effect is more clear (certainly when you group them together), will differ at the same lines/borders as dominant payment system will differ. If we want to isolate the dominant payment effect, the only reliable technique to do so would be meta analysis with a distinction of subgroups (with additional issues of clinical heterogeneity in other aspects and the occurrence of too few numbers of studies in subgroups if we want to make them sufficiently homogeneous). Both the reason of lack of information and the need for a more rigorous analysis have led us to conclude that this will not be feasible within this review paper. The paper gives an overview, while such an approach would go into depth for one characteristic. You are right in saying that we do have a large set of data available from this review to proceed further in this direction, but we think the issue is so complex that it should be dealt with in a separate study and paper.

b. Differences in results based on type of organization (primary care vs. hospital care) are presented as part of the study results, more specifically with regard to clinical effectiveness effects. Preventive and chronic care (see table 1) represent primary care; acute care represents hospital care. Of course in reality this distinction is erroneous (GPs also provide acute care, etc.), but for our study it fits remarkably well (because P4P programs do focus on those targets/indicators which are really typical for each setting; independent from any judgment from our part whether that’s a good thing or not). So, reporting effects by organization type would be almost identical to the current grouping into preventive/acute/chronic.

2. We agree there was some confusion possible regarding the meaning/definition of ‘outcome’, ‘mechanism’ and ‘context’. Outcome has therefore been replaced with ‘effects’ and mechanisms with ‘development choices’. Further clarifications have been provided. See also above, paragraph 1.1, 2.1.1, 5.1 and 6.1 in reply to Prof. Eijkenaar comments + subsequent changes.

3. This is another interesting suggestion. Our analysis however shows that it certainly is not a clear-cut relationship across studies, evolving through time toward larger sets. The results remain mixed and highly variable, also when grouped based on number of targets. As was the case for the dominant payment system analysis, meta analysis could maybe provide further insights in this, but it falls outside the scope of the current paper.
4. We added a theoretical introductory paragraph to reorganize provider characteristics in a logical manner. The target level of P4P incentives has been moved to the P4P incentive section as suggested.

5. This has been thoroughly revised in the abstract, the objectives and the discussion. See above, paragraph 1.2, 2.1, 5.1, 6.1, 7.15 in reply to Prof. Eijkenaar comments + subsequent changes.

6. Absence vs. conflicting evidence with regard to these recommendations has been further specified. The dose response relationship is very conflicting when analyzed across studies (small with very positive effects, large with lesser effects, ...but sometimes also the expected pattern). This implies other factors mediate its role. See previous remarks about a need for meta analysis by subgroups to proceed further (outside the scope of this paper).

7. We added the limitations regarding ceiling due to voluntary participation, publication bias and the distorting effect of co interventions, as suggested.

8. For reasons of fit with another upcoming paper, presenting a comprehensive conceptual framework, we decided not to integrate ‘evaluation’ into the previous paragraph. Else there also might be a loss of structure and coherence within the paper.

9. Publications in years 2000-2006: two in 2000, two in 2001, one in 2002, six in 2003, seven in 2004, five in 2005, twenty one in 2006. Readers can find this information in eTable 8. We think the trend is accurately covered by the sentence ‘since 2000 this number increased to more than 20 studies…’

10. The same finding is observed in the UK, where all CHD patients in primary care (on a national level) are part of the target population. This makes such a hypothesis less likely, although this might be health system specific. For example, a large number of quality initiatives preceded P4P aimed at CHD patients in the UK. This decreased room for improvement in comparison with other patient groups. Maybe other factors apply for the US, which are currently still unclear. Or do you refer to a diluting effect, which would be more typical for the US?

11. a. Clarified and expanded
b. Expanded
c. Deleted part of the sentence for clarification

12. Intermediary outcome measures like patients’ blood pressure value and HbA1c values for diabetes patients are less controllable for providers than process measures (e.g.
measuring blood pressure, HbA1c as an activity) and more controllable than long term outcomes (complication rate, mortality). This is a subjective interpretation and is therefore not added to the result section.

13. Acute care in terms of patient groups is presented before, in Table 1.

14. This is impossible to specify further for all studies within the paper’s length beyond what is covered under (3) (incentive size/amount is described). We do provide an additional excel format evidence table as supplementary information (see 3.7 in reply to Prof. Eijkenaar above).

15. See reply to your comment #6 above

16. See reply to your comment #1 above. You are right in saying that there is a lot of variation across countries. However, these differences make country comparison more instead of less interesting, as long as you always relate each result to the country where it is found. Our review implies that the success for the UK has to do with a combination of factors: incentive size, positive nature, high involvement and support (democratic decision process by the providers themselves), a uniform approach without dilution, a non tournament approach, etc. These factors are well aligned. The US does in general not worse, but more variable, because all kind of P4P configurations are tried with sometimes better and sometimes worse results. One program, the Premier hospital program, does combine a number of less supported aspects (low incentive of 1-2%, tournament,…). Despite its lower results in most studies (contrary to their own internal findings initially), it is still being propagated as a key role model for further dissemination (e.g. in one region in the UK a similar set-up is now being tested, with disappointing results). We understand the practical reasons for their type of set-up (minimal costs and maximal budget safety), but our review says – cautiously – that this might have its consequences in terms of effectiveness.

17. This theoretically based recommendation refers to the chicken or egg discussion, which comes first (I don’t know whether this is a regular expression in English, but you’ll get the grasp). Many authors argue that if you start with financial incentives, you should also provide the means to participants to change their behavior, change practice (additional tools and resources, e.g. staff) (external support). Other authors suggest that when you start with an incentive, the additional tools and resources (staff) will automatically follow as a consequence of the internal drive to increase performance. Right now multiple studies in the UK (qualitative nature, not included in this review) support that both will occur. External support will help providers getting started with improvement, and later on
they themselves redirect investments toward quality (tools, staff increase) as an internal mechanism. As mentioned in the section about P4P implementation in the paper, there is evidence support for the benefit of increased support in all countries except mixed findings in the US (due to its inherent higher variability). To be cautious we only mention it as a theoretically based recommendation.

18. You can indeed argue about the positioning of patient behavior. We feel it is better to include it under patient characteristics, because trying to incorporate it anywhere else led to a less cohesive structure.

19. We added a paragraph to the discussion (after recommendations, before limitations).

Reviewer Anne Frolich

1. Significance of results is reported in Table 1 (n.s. as not significant versus everything else which is significant; the amount of studies does not allow to specify all significance levels further). However, we agree that the strength of results had to be clarified further. Therefore we added more research design information when reporting results on the various quality domains (e.g. additional sentence on strength of cost effectiveness results). Throughout the text we also reformulated a number of statements more cautiously, e.g. low instead of fair evidence). See paragraph 1.2, 4.7, 5.16, 7.3, 7.7 in reply to Prof. Eijkenaar above. Furthermore, quality appraisal results can be consulted in the supplementary material.

2. We clarified ‘positive’: 1-10% and ‘very positive’: above 10%. Of course such intervals are based on a subjective appreciation, with differing opinions.

3. These targets are presented in Table 1. We added ‘preventive care’ into the table to make the distinctions more clear.

4. Mechanism has been replaced with ‘development choices’ throughout the text.

5. This has been specified: ‘chart audit, claims data, newly collected data’

6. References added as suggested

7. These important findings are described in the equity section previously.

8. These are theoretical recommendations to give further guidance where evidence is absent or conflicting. A complete theoretical framework cannot be added to the paper due to restrictions of its length. We do refer to some other papers, and another paper on a framework is submitted elsewhere.