

**What impact does physician pay-for-performance program  
have on the quality of health care in office-based chronic  
disease care? – A systematic review**

Submitted by:

Mr. Yong Shi  
Peace Country Health, Alberta

Advisor: Dr. Trish Reay  
SEARCH Faculty  
Assistant Professor, University of Alberta Business School

For

SEARCH Canada

December 2008

## Table of Contents

Executive Summary .....	1
Introduction.....	3
Scope of the Study .....	4
Pay-for-performance (P4P) .....	5
Quality of care .....	5
Chronic disease care.....	5
Literature Search Strategies.....	6
Databases .....	6
Key words used.....	6
Inclusion criteria .....	7
Exclusion criteria .....	7
Results .....	7
Intervention (i.e. P4P).....	8
Funding source .....	8
Funding size.....	8
Funding recipient .....	8
Types of Quality Measures.....	9
Dimension of Quality Indicators .....	9
High performance vs. Quality improvement .....	9
Outcome (i.e. effects on quality measures).....	9
Adverse Effects (i.e. unintended consequences).....	10
Discussion.....	10
Conclusion .....	14
Acknowledgement .....	15
References .....	16
Appendix A. Literature Search Algorithm.....	18
Appendix B. Methodological Quality of Included Articles .....	19
Quality Index scores – a brief description .....	19
Quality Analysis of Included Articles.....	19
Appendix C. Study Assessment.....	21

## **Executive Summary**

The urgent need to improve performance in health care is now well recognized worldwide, and there is a new trend to use financial incentives to meet this need.

Various types of Pay-for-performance (P4P), an incentive program, are getting popular in a few countries including U.K., U.S., Australia, and Canada. One consistent theme for these incentive programs is the need to better align the incentives of providers with the health system goal that is to form a new service delivery model by leveraging incentives to improve quality of care and patient safety. However, the effectiveness of these programs has not been thoroughly evaluated.

The purpose of this paper is, by systematically reviewing studies in U.K. and U.S., to assess the relationship between explicit financial incentives and the quality of health care in chronic disease care, and to identify facilitators and barriers in the design and implementation of the P4P program.

A systematic search of English-language, peer-reviewed literatures was conducted to review articles published from January 2000 to March 2008. Study eligibility was determined in accordance with inclusion and exclusion criteria. Six articles were included in the final analysis: 4 Cohort studies and 2 Controlled pre/post studies. Three factors in each article were assessed: the intervention (i.e. P4P), types of quality measures, and the outcome (i.e. effects on quality measures).

These six studies showed various degrees of positive association between a P4P program and the improvement in the quality of care. However, for several reasons, the generalization from the existing studies was limited. First of all, from a methodological perspective, the design of these studies was not rigorous. Secondly, these studies

focused on a narrow set of quality indicators. Finally, due to the nature of American service delivery and insurance systems, the study conclusion from one type of organization was hard to be generalized to represent other types of organizations.

Despite these limitations, a few preliminary conclusions could be drawn: 1) P4P program requires careful design of financial rewarding, including funding source, size, and recipient; 2) P4P program requires careful selection of quality indicators and performance targets; and 3) P4P program needs an integrated database.

There are still many questions to be answered and many myths to be revealed. Further research is required to understand the most appropriate approach to address both issues of quality improvement and healthcare cost constraints.

This review may be beneficial for the potential consideration of a similar program in Alberta's primary care settings. Meanwhile, Primary Care Network (PCN) initiative itself in Alberta provides a great opportunity for piloting a P4P program and testing its impacts on primary care.

## **Introduction**

The urgent need to improve performance in health care is now well recognized worldwide, and there is a new trend to use financial incentives to meet this need. In 2004, National Health Service (NHS) in U.K. committed \$3.2 billion in additional funding over a period of three years for a new pay-for-performance (P4P) program for family practitioners, of which 146 quality indicators are applied (Doran et al., 2006). A recently published paper reports that over 100 quality incentive programs now exist in U.S. in both private and public sectors (Baker and Carter, 2005). In addition, the P4P program has attracted considerable interest in other countries: in Australia, Medicare's Practice Incentive Program targeting quality in general practice is being piloted from July 2007 (Ward, Daniels, Walker, and Duckett, 2007); and in Canada, certain variations of such incentive programs could be recently found in Ontario and Saskatchewan, though they may have different names. However it is still too early to evaluate how these Made-in-Canada programs work and what impacts they have on the primary care delivery system.

One consistent theme for these incentive programs is the need to better align the incentives of providers with health system goals (Custers, Hurley, Klazinga, and Brown, 2008). A health system goal is to form a new service delivery model by leveraging incentives to improve quality of care and patient safety. Concerns about the quality of care are growing across the world. First of all, many studies indicated major deficits in provider compliance with evidence-based clinic practice (McGlynn et al., 2003). Secondly, "... lack of quality-related financial incentives in traditional provider reimbursement system is a major barrier to quality improvement in the United States"

(Institute of Medicine, 2001). Furthermore, Scott (2007) has made similar conclusions and found that traditional approaches to optimize care, including continuing medical education and certification, do not guarantee minimum standards. All these findings could contribute to the gradual popularity of P4P in U.K. and U.S., and recently in Canada and Australia.

However, the effectiveness of these programs has not been thoroughly evaluated. The purpose of this paper is, by systematically reviewing studies in U.K. and U.S., to assess the relationship between explicit financial incentives and the quality of health care in chronic disease care, and to identify facilitators and barriers in the design and implementation of a P4P program. In addition, following areas would also be investigated: the source, size and recipient of financial incentives, the selection of performance measures and targets, data collection methods, and the appropriate mix of financial and non-financial incentives. Given that P4P-related discussion has begun in this province, this review may provide in-depth understandings and implications to primary care renewal initiatives in Alberta.

## **Scope of the Study**

Due to limitations of time and resources, as well as the relevance to my work environment, this paper is limited to evaluating primary care physician P4P, within the range of chronic disease care for out-patients. Any P4P programs designed for physicians other than primary care physicians, for acute care and in-patient care, and for hospitals are excluded. Thus, in order to clarify the scope of this review, following terms used throughout this paper need to be defined.

### **Pay-for-performance (P4P)**

A form of reimbursement based on providing explicit financial incentives to physicians for attaining certain quality benchmarks established by the payers (Spinelli and Fromknecht, 2007). This new form is so different from Fee-for-service, the dominating fee schedule in primary care in Alberta, that the former one links reimbursement to “quality” of services, and the latter one is actually “pay-for-volume”, i.e. reimbursement linked to volumes of cases and procedures.

### **Quality of care**

Lohr (1990) defined the quality of care as “ the degree to which health care services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge”. Donabedian (1988) used “structure, process, and outcome” taxonomy for targeting dimensions of quality, all of which will be assessed in this paper.

- Structure: personnel, facilities, and materials to assist in physicians and clinics.
- Process: the completion of specific tasks or recommended treatments. Clinic practice guidelines (CPGs) do help to define standards of care and focus efforts to improve quality. Quality of care sometimes is considered as “physician adherence to interventions reflecting CPG recommendations” (Boyd et al., 2005).
- Outcome: the ultimate results of care, including patient’s experience and health status. It includes immediate, intermediate, and long-term outcomes.

### **Chronic disease care**

Chronic disease care is defined as office-based medical care provided to patients with chronic disease(s), including cardiovascular diseases, respiratory illness, diabetes,

P4P and Quality of care

arthritis, and mental health issues. In addition, screening for some types of cancer (i.e. cervical cancer, breast cancer and colorectal cancer) is also considered chronic disease care. A 2003 report estimated that almost 60 percent of Canadians 12 years of age or older have at least one chronic disease (Schultz and Kopec, 2003). The total costs associated with caring for individuals with chronic disease in Canada are estimated to exceed \$80 billion annually (Chronic Disease Prevention Alliance of Canada, 2006). The need for appropriate chronic disease care has posed great challenges on the healthcare system.

## **Literature Search Strategies**

A systematic search of English-language, peer-reviewed literatures was conducted in following databases to review articles published from January 2000 to March 2008, with a main objective to assess the relationship between explicit financial incentives and quality of care. In addition, related publications found in bibliographies of retrieved articles were also reviewed. Study eligibility was determined in accordance with following inclusion and exclusion criteria.

### **Databases**

PubMed, ABI, EBSCO (Health Business Fulltext Elite, Psychology and Behavioural Science Collection, Nursing & Allied Health Collection: Comprehensive, Biomedical Reference Collection: Comprehensive, CINAHL Plus with Full text, Medline with Full text, and Medline), and all EMB Reviews (Cochrane DSR, ACP Journal Club, DARE, CCTR, CMR, HTA, and NHSEED).

### **Key words used**

Included: quality of care, pay for performance, primary care, physician.

### **Inclusion criteria**

- Independent variable: explicit financial incentives;
- Dependent variable: measures of quality of care; and
- Controlled studies including RCT or Observational studies including cohort studies or cross-sectional studies.

### **Exclusion criteria**

- Not for office-based chronic disease care; and
- No pre-intervention baseline of quality indicators and/or a comparison or controlled group.

## **Results**

The search algorithm for eligible studies is exhibited in **Appendix A**. Six articles were included in the final analysis: 4 Cohort studies and 2 Controlled pre/post studies.

Excluded studies included one (Felt-Lisk, Gimm, and Peterson, 2007) assessing the incentive effect on the quality of well-baby care, and the other (Levin-Scherz, DeVita, Timbie, 2006) lacking a baseline during pre-P4P implementation or a comparison group after the P4P program being implemented.

A checklist for methodological quality published by Downs and Black (1998) was used to assess the quality of each individual article. A grade of 1 (poor) to 4 (excellent) was assigned to each included article (**Appendix B**). All 6 articles were graded at 3 (good); and their Global scores ranged from 18-23.

**Appendix C** summarized the details of 6 included studies. Three factors in each article were assessed: the intervention (i.e. P4P), types of quality measures, and the outcome (i.e. effects on quality measures). In addition, adverse effects or unintended

P4P and Quality of care

consequences, if there were any, would be identified as cautions for the potential consideration of P4P program design and implementation.

### **Intervention (i.e. P4P)**

#### *Funding source*

Five of six articles indicated that their individual payers committed additional funding to support this program; in other words, physicians were not at financial risks that their base incomes would be impacted by the implementation of P4P programs. Only one article (Young et al., 2007) indicated a “withhold” approach was used for the P4P program. The “withhold” put physicians at financial risks, which required each individual physician to contribute 5% of physician fees to fund the incentive pool.

#### *Funding size*

NHS used a point system for financial rewarding across the country. The maximum P4P funds a physician may get reimbursed accounted for up to 25% of her/his annual income. In U.S., due to various organizational types of providers including Health Maintenance Organization (HMO), Preferred Provider Organization (PPO) and Managed Care Organization (MCO), and various physician payment schedules including capitation, salary, and fee-for-service, P4P funding size varied from one organization to another: from the lowest 1% to the highest 7.5% of annual incomes.

#### *Funding recipient*

Except one program (Rosenthal et al., 2005) which financially rewarded individual physician groups instead of individual physicians, the remaining five all reimbursed individual physicians directly. Some uniqueness existed with the program rewarding physician group: physician groups paid capitation and performance measured at physician group level.

## **Types of Quality Measures**

### *Dimension of Quality Indicators*

Three programs (Rosenthal et al. 2005, Gilmore et al. 2007, and Young et al. 2007) used only “process” measures for the designing and rewarding system. The common indicators included HbA1c testing, LDL screening, cervical cancer screening, and mammography. One article (Millett et al. 2007) indicated only “immediate outcome” indicators used in the system, including the control of HbA1c, BP, and total cholesterol levels. The other two programs (Campbell et al. 2007, and Beaulieu & Horrigan 2005) adopted a mix system for measuring performance, a combination of process and immediate outcome indicators.

### *High performance vs. Quality improvement*

NHS and one program affiliated with a HMO (Rosenthal et al.) predetermined a set of indicator thresholds or benchmarks as the only criterion for the consideration of rewarding. One MCO program (Beaulieu & Horrigan) used a combination approach: one physician receiving 100% of incentives for reaching the “best” benchmarks, or 50% for meeting the second to the “best” benchmarks, or 25% for unable reaching either of these two benchmarks but achieving 50% improvement from their individual baselines. The other two programs (Gilmore et al. & Young et al) used “ranking” system: neither threshold nor improvement range was predetermined; each individual physician would be ranked among all participating physicians and reimbursed accordingly.

### **Outcome (i.e. effects on quality measures)**

In general, all P4P programs saw improvement in quality of care when linked to the financial incentives. However, the extent to how well the P4P impacted the quality improvement varied from one study to others.

## P4P and Quality of care

Both of two NHS Cohort studies indicated a significant association between the P4P implementation and the achievement of quality measures over the study period; but the P4P was unable to eliminate the ethnic disparities already existing.

One PPO Cohort study (Gilmore et al.) found a positive association between the quality of care delivered to the patients and whether they received care from a participating physician or not. And another HMO cohort study (Young et al.) also made such similar conclusion; however, this 6-year study also revealed an important phenomenon that quality improvement in post-P4P trends was not different from that in pre-P4P trends.

Both of 2 Controlled studies (Beaulieu & Horrigan, and Rosenthal et al.) found that within the study group the P4P programs had made a significant, positive impact to improve the quality of care, from year to year. With respect to the comparison between the study group and the control group, Beaulieu & Horrigan indicated a significant difference, but not the same case for the study reported by Rosenthal et al., which found only 1 of 3 indicators made the difference-in-differences.

### **Adverse Effects (i.e. unintended consequences)**

No adverse effect was explicitly reported in the studies. Only one article reported a possible unintended consequence. Campbell et al. indicated their results that the program assessment was based on the medical results, which didn't necessarily reflect the actual services provided to patients, implying physicians "gaming" with the rewarding system may exist.

## **Discussion**

Six articles were found to be eligible in this systematic review, which partially addressed the question of "what impact does physician pay-for-performance program have on the

quality of health care in office-based chronic disease care ?” Two Cohort studies assessed P4P programs within NHS in U.K., and the other four were in the U.S.: two with HMO, one with PPO, and the forth with MCO. Among these four articles in U.S., two used a Cohort study design , and the others adopted a Controlled pre/post study design. These six studies showed various degrees of positive association between the P4P program and the improvement in the quality of care. However, for several reasons, the generalization from the existing studies was limited.

First of all, from a methodological perspective, the design of these studies was not rigorous. The quality of all six studies was graded at 3 (“Good”), not “Excellent”. Data in two NHS studies was collected only at 2-3 time points, and only one of these time points was during post-intervention, which led to questioning the sustainability of the positive impact. In addition, because of the universal implementation of the P4P program within NHS, reasons why those physicians who didn’t achieve the predetermined targets were not described. For US studies, selection bias could not be excluded. Since the participation in P4P programs was voluntary, and most programs received additional funding, a possibility existed that the participating physicians were already higher performers and knew they would get extra dollars. Though there was a “controlled” group in the two controlled studies respectively, how well these groups were controlled were not described, and how confounders, e.g. fee schedules, patient demographics, insurance status, etc., were compared were also not articulated. Therefore, the secular trend of quality improvement was not possible to be excluded.

Secondly, these studies focused on a narrow set of quality indicators. For instance, 3 of 6 studies only targeted “diabetes”. Though the study results showed positive impacts, the generalization to other chronic diseases including COPD, CAD, and mental health

would be questionable, since physicians may have different practice behaviours, as well as patients' demographics and behaviours.

Finally, due to the nature of American service delivery and insurance systems, the conclusion from a PPO is hard to be generalized to represent other organizations like HMOs. Even among the same type organization like HMO, different payers and different insurance plans would make such generalization difficult.

Despite these limitations, a few preliminary conclusions could be drawn.

- 1) P4P program requires careful design of financial rewarding, including funding source, size, and recipient. NHS committed a very large amount of money over a 3-year period of time to support the program. Individual physicians could make up to 25% of their annual incomes if achieving full scores. It should be cautioned that how sustainable this program would be in terms of the issue of healthcare cost constraints and ongoing financial support, and that what significant negative impacts may emerge if NHS would withdraw the program at the end of the 3-year period. Several American programs allocated the equivalent of 1%-7.5% of physician's average annual income as the incentive. One percent probably was not attractive enough for a physician to change the status quo, which might explain why some programs only made small effects.

In addition to the source and the size, who eventually received financial incentives may play an important role in impacting program effects. Most studies (i.e. 5 of 6) collected data on individual physician level and rewarded accordingly, which greatly reduced the potential possibility of low-performers' free-riding on efforts of others in such program that a physician group was the receiver.

2) P4P program requires careful selection of quality indicators and performance targets. "Process" indicators like the frequency of HbA1c testing, blood pressure measuring, and consultation provided to CHD patients, were more acceptable by physicians and feasible for implementation, and could be easily collected through administrative or claim databases. However, these process improvements may not necessarily lead to a better health status or outcome. On the other hand, "outcome" measures including the control of HbA1c, LDL, and BP were more meaningful to payers and more beneficial to patients. But the measurement and comparison of these indicators might face resistance from physicians, since context factors including patient mix, disease severity, access to other services, etc. may be very different from one physician to others, and from one geographic location to others. These factors must be considered and adjusted to minimize the "unfairness" in order to improve the credibility and acceptance of a P4P program.

Additionally, P4P program designers were facing another challenge: what type of performance targets to be set for rewarding: high performance or performance improvement. Both of these two types had pros and cons. The former one recognized existing high performers, but who may lack motivation to continue their efforts in improving quality; meanwhile, this method discouraged the motivation of those low performers who may feel the target was set too high for them to achieve, and as a result they may opt out of this program. The latter type did encourage and reward low performers for their efforts and achievement, but might penalize high performers because of the "ceiling effect", small room and difficulty for high performers to continue improving the already high performance. Taken together, a well designed P4P program

should include a careful combination of both process and outcome indicators, and target on both high performance and performance improvement.

3) P4P needs an integrated database. Most studies used administrative data for quality analysis and comparison. Administrative data was appropriate and worked well for collecting “process” measures, i.e. what and how many services were provided to a patient. But a medical record would tell us the real story of that patient, i.e. how well that patient was treated and what the health status he/she is presently at, which is the main intent for a payer to consider and adopt such P4P program. Also, fragmented databases in the American environment limited the P4P program’s application and comparison across sections. An integrated, universal database that could accommodate both of these two needs (i.e. process and outcome indicators) would make it possible for a feasible, measurable, accountable, and credible P4P program.

## **Conclusion**

This systematic review provided an in-depth understanding of the positive association between P4P programs and quality of care in chronic disease care in U.K. and U.S. In addition, the review described and analyzed several major components in P4P program design, including funding source, size, and recipient, and performance measures regarding indicators and targets, as well as a few limitations within the existing studies. There are still many questions to be answered and many myths to be revealed. How a P4P program may impact other types of practice settings including hospital-based care and interdisciplinary team-based care? How non-financial incentives including IT, chart auditing and feedback, and administrative support, may assist in the quality improvement? Is a P4P program cost-effective? And so on. Further research is required

P4P and Quality of care

to understand the most appropriate approach to address both issues of quality improvement and healthcare cost constraints.

This review may be beneficial for the potential consideration of a similar program in Alberta's primary care settings. Meantime, Primary Care Network (PCN) initiative itself in Alberta provides a great opportunity for piloting a P4P program and testing its impacts on primary care, for the following two reasons: 1) a universal payer and committed additional funding; and 2) certain levels of infrastructures already build up including IT/IM integration, administrative support, regular data collection, organized physician groups, regular reporting, and existing programs of inexplicit financial incentives on the quality of care.

## **Acknowledgement**

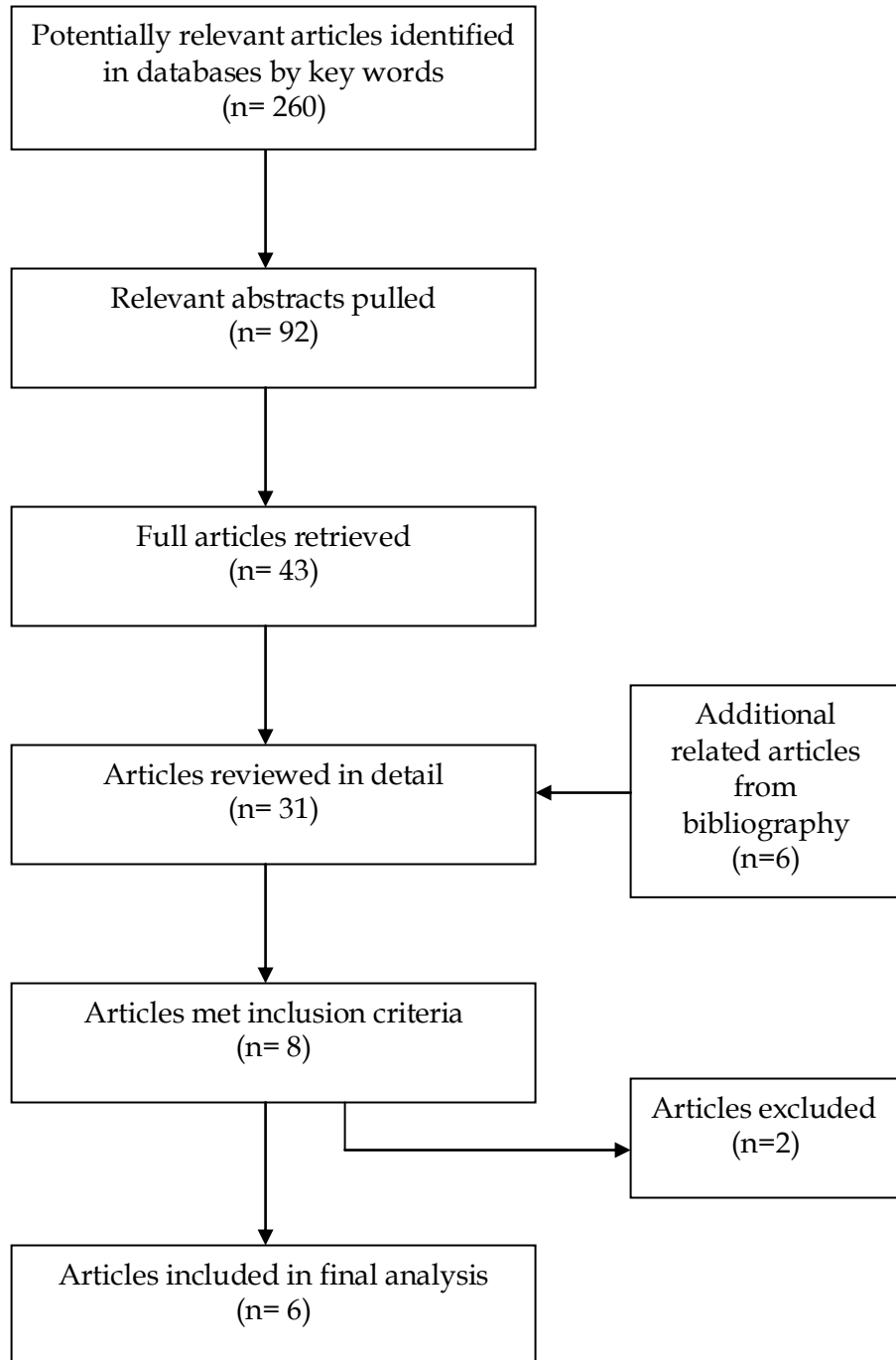
Thank SEARCH Canada and Peace Country Health for their kindly support in providing funding and organizing SEARCH VI study. Also thank Dr. Trish Reay, SEARCH faculty, and Mr. Sean Chilton and Ms. Sherryl Hoskins from Peace Country Health for their contribution of their valuable time, advice and consultation, and for their ongoing commitment in supporting my participation in SEARCH VI.

## References

- Baker, G., and Carter, B. Provider pay-for-performance incentive programs: 2004 National Study Results. *San Francisco: Med-Vantage* 2005.
- Beaulieu, N., and Horrigan, D. Putting smart money to work for quality improvement. *Health Services Research* 2005; 40 (5): 1318-34.
- Boyd, C., Darer, J., Boulton, C., Fried, L., Boulton, L., and Wu, A. Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. *JAMA* 2005; 194 (6): 716-24.
- Campbell, S., Reeves, D., Kontopantelis, E., Middleton, E., Sibbald, B., and Roland, M. Quality of primary care in England with the introduction of pay for performance. *N ENGL J MED* 2007; 357 (2): 181-89.
- Chronic Disease Prevention Alliance of Canada (CDPAC). *The Case for Change*. Retrieved on January 19, 2006:  
[http://www.cdpac.ca/content/case\\_for\\_change/case\\_for\\_change.asp](http://www.cdpac.ca/content/case_for_change/case_for_change.asp).
- Custers, T., Hurley, J., Klazinga, N., and Brown, A. Selecting effective incentive structures in health care: a decision framework to support health care purchasers in finding the right incentives to drive performance. *BMC Health Services Research* 2008; 8: 66.
- Donabedian, A. The quality of care. How can it be assessed? *JAMA* 1998; 260 (12): 1734-78.
- Doran, T., Fullwood, C., Gravelle, H., Reeves, D., Kontopantelis, E., Hiroeh, U., and Roland, M. Pay-for-performance program in family practices in the United Kingdom. *N ENGL J MED* 2006; 355: 375-84.
- Down, S, Black, N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomized and non-randomized studies of health care interventions. *J Epidemiol Community Health* 1998; 52: 377-84.
- Felt-Lisk, S., Gimm, G., and Peterson, S. making pay-for-performance work in Medicaid. *Health Affairs* 2007; 26 (4): w516-27.
- Gilmore, A., Zhao, Y., Kang, N., Ryskina, K., Legorrela, A., Taira, D., and Chung, R. Patient outcomes and evidence-based medicine in a Preferred Provider Organization setting: a six-year evaluation of a physicians pay-for-performance program. *Health Services Research* 2007; 42 (6): 2140-59.
- Institute of Medicine. *Crossing the quality chasm: a new health system for the 21<sup>st</sup> century*. Washington, DC: National Academies Press 2001.
- Levin-Scherz, J., DeVita, N., Timbie, J. Impact of pay-for-performance contracts and Network Registry on diabetes and asthma HEDIS Measures in an integrated delivery network. *Medical Care and Review* 2006; 63 (1): s14-s28.
- Lohr, K. *Medicare, a strategy for quality assurance*. Washington, DC: National Academy Press 1990.

- Millett, C., Gray, J., Saxena, S., Netuveli, G., Khunti, K., Majeed, A. Ethnic disparities in diabetes management and pay-for-performance in the UK: the Wandsworth prospective diabetes study. *PloS Medicine* 2007; 4 (6): 1087-93.
- McGlynn, E., Asch, S., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., and Kerr, E. the quality of health care delivered to adults in the United States. *The New England Journal of Medicine* 2003; 348 (26): 2635-45.
- Rosenthal, M., Fernandopulle, R., Song, H., and Landon, B. Paying for quality: providers' incentives for quality improvement. *Health Affairs* 2004; 23 (2): 127-41.
- Rosenthal, M., Frank, R., Li, Z., Epstein, A. Early experience with pay-for-performance: from concept to practice. *JAMA* 2005; 294 (14): 1788-93.
- Schultz, S., Kopec, J. Impact of chronic conditions. *Health Reports* 2003; 14(4): 41-53.
- Scott, I. Pay for performance in health care: strategic issues for Australian experiments. *MJA* 2007; 187 (1): 31-35.
- Spinelli, R., and Fromknecht, J. Pay for performance improving quality care. *The Health Care Manager* 2007; 26 (2): 128-37.
- Ward, M., Daniels, SA., Walker, GJ., and Duckett, S. Connecting funds with quality outcomes in health care: a blueprint for a Clinical Practice Improvement Payment. *Aust Health Rev* 2007; 31 (Suppl 1): s54-s58.
- Young, G., Meterko, M., Beckman, H., Baker, E., White, B., Sautter, K., Greene, R., Curtin, K., Bokhour, B., Berlowitz, D., and Burgess, J. Effects of paying physicians based on their relative performance for quality. *Society of General Internal Medicine* 2007; 22: 872-76.

## Appendix A. Literature Search Algorithm



## Appendix B. Methodological Quality of Included Articles

### Quality Index scores - a brief description

Downs and Black (1998) developed a checklist claimed to be capable of assessing the methodological quality not only of RCTs but also of non-randomized studies. Due to natures of the studies evaluated, none of 6 included articles was RCT; thus it might be appropriate to adopt this checklist for the quality assessment. The checklist, Quality Index with 5 sub-scales, is briefly described below:

Sub-scales	No. of Items	Max. Scores
Reporting	10	11
External validity (Generalizability)	3	3
Internal validity (Bias)	7	7
Confounding (Select bias)	6	6
Power	1	5
<b>Global</b>	<b>27</b>	<b>32</b>

However, Downs and Black didn't provide any methodology in the article to link the Global scores to the grade of "quality". In this paper, the "quality" is stratified into 4 grades, from 1 (poor) to 4 (excellent), simply on the basis of the Globe Score quartile.

Global Scores	Quality Grade
1-8	1 (poor)
9-16	2 (satisfied)
17-24	3 (good)
25-32	4 (excellent)

### Quality Analysis of Included Articles

Study	Quality Index scores						Grade	Weaknesses
	R	E	I	C	P	Global		
Campbell et al. (2007)	8	3	4	2	5	22	3	Only treated patients included; and no control group due to the single-payer system
Millett et al. (2007)	9	1	4	2	5	21	3	The findings may not be generalized to the entire population in UK; no control group due to the single-payer system; and some confounders not adjusted.

Study	Quality Index scores						Grade	Weaknesses
	R	E	I	C	P	Global		
Beaulieu & Horrigan (2005)	8	0	5	2	5	20	3	How well the selected physicians represent the entire physician community is not described; and final results may not be generalized to other physician groups.
Rosenthal, Frank, Li, Epstein (2005)	8	2	5	3	5	23	3	Physician groups studied were compensated by capitation, which limited its generalizability to other settings, in particular those paid by fee-for-service.
Gilmore et al. (2007)	5	0	5	3	5	18	3	Due to the nature of PPO, patients are free to visit any discipline of physicians within the organization. The study is unable to differentiate the P4P effect between primary care physicians and specialists.
Young et al. (2007)	8	1	5	2	5	21	3	The study was conducted in a limited area, the results may not be generalized to other settings.

## Appendix C. Study Assessment

Note: NHS=National Health Service; MCO=Managed Care Organization; HMO=Health Maintenance Organization; PPO=Preferred Provider Organization

Study (Year)	Org. Type	Study Design	Intervention (P4P)	Quality Measures	Outcomes
Campbell et al. (2007)	NHS	Cohort study  42 practices in 6 geographic areas of England.  P4P program implemented in 2004, and data available for all 3 time points (1998, 2003, & 2005).	Source: additional funding  Size: up to 25% of GP's income  Recipient: individual physicians	Indicators: CHD (15), asthma (12), type II diabetes (21).  Dimension: a mix of process and immediate outcome.  Only rewarding whose performance achieving predetermined thresholds	P4P was associated with a modest acceleration in improvement for 2 of 3 conditions: asthma & diabetes.  However, the result is based on medical records, but not necessarily on care provided.
Millett et al. (2007)	NHS	Cohort study  4,284 adults with diabetes with 32 practices in Wandsworth PCT, in which 22% of residents belongs to a non white	Source: additional funding  Size: up to 25% of GP's income  Recipient: individual	Indicators: diabetes (3)  Dimension: immediate outcome (HbA1c, BP, & total cholesterol)  Only rewarding whose performance achieving	The proportion of patients reaching thresholds of 3 indicators increased significantly after the implementation of P4P in 2004.  However, improvement in black Caribbean was

Study (Year)	Org. Type	Study Design	Intervention (P4P)	Quality Measures	Outcomes
		ethnic group.  P4P implemented in 2004, and data collected in both 2003 & 2005.	physicians	predetermined thresholds	significantly lower than that in White British group.
Beaulieu & Horrigan (2005)	MCO	Controlled pre/post study  Study group - 21 physicians with 624 diabetic patients, within Independent Health (IH) in upstate NY.  Control group - a sample of 600 of IHs diabetic members.  Data collected 3 times during one year (baseline, interim, & final).	Source: additional funding  Size: \$3,000 to \$12,000/year, based on panel size and performance  Recipient: individual physicians	Indicators: diabetes (9)  Dimension: a mix of process and immediate outcome.  Two different ways for rewarding: 1) achieving 1 of 2 benchmarks, receiving 100% or 50% of incentives, or 2) 50% improvement, receiving 25% of incentives	Significant improvement on 5 out of 6 process measures both in the study group and between the study and control groups. The performance of the study group at baseline is in all case below the baseline of the control group.  Outcome measures including HbA1c and LDL also increased significantly both in the study group and between the study and control groups.  BP control improvement among the study group was modest but still significant, but lack of such data in the control group.

Study (Year)	Org. Type	Study Design	Intervention (P4P)	Quality Measures	Outcomes
Rosenthal, Frank, Li, Epstein (2005)	HMO	Controlled pre/post study  Study group – 163 California physician groups.  Control group – 42 Pacific Northwest physician groups.  Performance in 2002 set as the baseline. Incentive program began in 2003. Data collected pre- and post-implementation.	Source: additional funding  Size: 5% of capitation & 0.8% of the groups' overall revenue  Recipient: individual physician groups	Indicators: 3  Dimension: process (the rate of cervical cancer screening, mammography, & HbA1c testing)  The 75 percentile of the baseline set as the performance target. Only rewarding whose performance achieving the targets.	In the study group, a significant improvement of cervical cancer screening occurred after P4P, and a modest improvement for the other two.  However, among the difference-in-differences, only the difference for cervical cancer screening between the study and control groups was significant (P=0.02), but no difference for the other two.
Gilmore et al. (2007)	PPO	Cohort study  A 6-year study in Hawaii Medical Services Association, which includes 95% of physician in Hawaii. The incentive program launched in 1998.	Source: additional funding  Size: 1-5% of base professional fees in 1998-2001, and increased to 1-7.5% in 2002-2003, average \$4,785.	Indicators: 11  Dimension: process (e.g. cancer screening, HbA1c testing, continuity of supply of drugs, etc.)  Physicians compensated by the calculation of percentile	A positive association between the quality of care delivered to patients and whether they received care from a incentive program-participating physician. Moreover, the association grew stronger in the later years when patients who visited only program-

Study (Year)	Org. Type	Study Design	Intervention (P4P)	Quality Measures	Outcomes
		Physician participation increased from 50.4% in 1998 to 77.7% in 2003.	Recipient: individual physicians	ranking individually among all participating physicians.	participating physicians were compared with members who visited only non-participating physicians.
		Data collected 1998-2003. The data analysis was controlled for any baseline temporal trends by including a local control group (non-participating physicians).			However, the study revealed a negative trend in the association between the quality of care and the group of patients who saw both participating and non-participating physicians.
Young et al. (2007)	HMO	Cohort study  A 6-year study involving 334 primary care physician in Rochester, NY.  The incentive program initiated in 2002, and performance data collected from 1999 to 2004 (3 years before and 3 years after).	Source: "Withhold"- each physician to contribute 5% of physician fees, about \$8,000-\$14,000 of annual income, to fund the incentive pool Size: receiving 50-150% of the contribution. The diabetes measures	Indicators: diabetes (4), (accounted for 50% of the total score for the quality component of the entire performance measure)  Dimension: process (e.g. LDL screening, HbA1c testing)  Physicians ranked annually according to the overall adherence score for each set of performance, then	A statistically significant increase in performance levels after the incentive program implementation. However, the post-intervention trends were not different from the pre-intervention trends.  When assessing year-to-year changes in performance scores, there was a significant increase in the eye examination performance score.

Study (Year)	Org. Type	Study Design	Intervention (P4P)	Quality Measures	Outcomes
		Data analysis was compared to national and NY trends over the study period.	<p>were worth approximately 1% of a typical PCP's earnings.</p> <p>Recipient: individual physicians</p>	compensated accordingly.	