Lieve opa, ik draag dit proefschrift aan u op.

Improving primary care by pay-for-performance

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus
prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op
donderdag 10 juli om 14.30 uur precies

door ***Kirsten Kirschner***

geboren te Nijmegen op 2 september 1981

Promotor	Prof. dr. R.P.T.M. Grol
Copromotoren	Dr. J.C.C. Braspenning
	Dr. J.E. Jacobs
Manuscriptcommissie	Prof. dr. J. van der Velden (voorzitter)
	Prof. dr. G.J. van der Wilt
	Prof. dr. R. Bal (Erasmus Universiteit Rotterdam)

Content

The studies presented in this thesis have been performed at the Scientific Institute for Quality of Healthcare (IQ healthcare).
This institute is part of the Radboud Institute for Health Sciences (former NCEBP), one of the approved research institutes of the Radboud University Medical Centre.
The studies in this thesis were supported by CZ and VGZ, the two main health insurance companies in the south of the Netherlands.

Zetwerk & opmaak Zetterij Chang Chi Lan-Ying
Drukwerk Offsetdrukkerij Jan de Jong
Papier Igepa Nederland
Afwerking Binderij Patist, Den Dolder
Oplage 400 exemplaren
Isbn 978 94 90913 43 4

Chapter 1

Introduction

This thesis concerns pay-for-performance (P4P) in primary care, more specific in general practice.
The introduction briefly describes the improvement of quality of care and gives an overview of the content and effectiveness of and the experiences with P4P programs. This overview will lead to a set of research questions that will be addressed.

Improvement of quality of care

The number of new insights, methods, programs and techniques that are available each year in healthcare is enormous. Healthcare providers have to deal with all these innovations in daily practice, but as has shown for instance by the introduction of guidelines, innovations do not implement themselves.[1] Effective improvement of the quality of patient care often starts with good data on the quality of care delivered. The aim of feedback is to present data about practice performance and to encourage practices to design specific practice-based improvement plans as a guide for change.[2] Feedback about individual performance compared with that of peer performance can be a powerful motivator for change.[3-6] To give caregivers feedback on their performance does not necessarily lead to behavioural changes, but feedback can be an important part of a multifaceted improvement program.[1] We studied whether stimulating practice-based improvement plans and information about best practices led to improvement in accessibility and availability in general practice. Practices received feedback with a benchmark of their peers accompanied with information of best practices. With this information they were stimulated to make practice-based improvement plans.
Feedback can be effective, but the effects are usually small to moderate.[1] We assume that achievement of optimal quality of care delivered by healthcare providers can be enhanced further by introducing an extrinsic stimulus, like pay-for-performance (P4P). Pay-for-performance is increasingly applied by payers in healthcare. We know that financial incentives for quality of care and improving quality of care might improve quality and efficiency.[7-9] The question is how exactly they work and what factors may determine their success.

Pay-for-performance programs

International interest in pay-for-performance (P4P) initiatives to improve quality of healthcare is growing. Worldwide many P4P programs exist. Most of these programs can be found in the UK and the USA, but also in Australia P4P programs have been implemented. A few examples will be presented below.

Quality and Outcomes Framework (QOF)

In 2004, the NHS in the United Kingdom implemented a primary care pay-for-

performance program as part of the Quality and Outcomes Framework (QOF), linking up to 25% of general practitioners' income to performance on 76 clinical quality indicators and a further 70 indicators relating to organisation of care and patient experience.[10] The clinical indicators mainly relate to processes – for example measuring disease parameters and giving treatment – with only 10 of the 76 original clinical indicators relating to intermediate outcomes. Practices earned points based on the proportion of eligible patients for whom the quality targets are achieved. GPs are permitted to use their clinical judgment to exclude inappropriate patients from achievement calculations, a process known as 'exception reporting'.

Integrated Healthcare Association P4P program

The Integrated Healthcare Association (IHA) P4P program is the largest non-governmental physician incentive program in the United States. The program provides physician groups with financial rewards based on their performance compared to quality and efficiency benchmarks. For Measurement Year (MY) 2012, there are four measurement domains: clinical quality (prevention, cardio-vascular, diabetes, maternity, musculoskeletal, and respiratory conditions), patient experience, meaningful use of health IT and appropriate resource use.[11]

Practice Incentive Program (PIP)

The Practice Incentive Program (PIP) was introduced in 2001 in general practices in Australia. The program includes 11 incentives including quality prescribing, diabetes, asthma, cervical cancer, indigenous health, e-health, after hours care, teaching, rural loading, aged care access, and a financial incentive aimed at insuring access to surgical, anesthetic, and obstetric services in rural area.[12]

The design of P4P programs

Unanswered questions about P4P programs are related to the optimal design, the effectiveness and the use of the P4P programs. P4P programs are very heterogeneous, but they share common features. Each P4P program is developed by discussing the objectives of the program, the performance measurement, the appraisal (unit of assessment, performance standards, analysis and interpretation of performance data) and reimbursement (financial rewards).[8,13-15] Table 1 presents the elements of a P4P program.

Current P4P programs are mostly designed and implemented top-down by policy makers and managers.[16] P4P programs can be seen as an innovation in care, and it is known that the sustainability of an innovation can be improved by involving target users.[17] It has also been suggested to involve target users in the develop-mental process of a P4P program, because this can contribute to the effect of incentivised indicators.[15,18] A more bottom-up procedure in designing a P4P

program may improve its future implementation and its effectiveness. Therefore it is important to develop the P4P program in a systematic way, using for instance a Delphi procedure.[19] The perspectives of all target users become distinct, and the decisions made are transparent for the target users. We explored such a bottom-up approach in this thesis.

Table 1 Elements of the P4P program

Component	Elements
Performance measurement	**Performance indicators** Domains, subjects and indicators Period of data collection
Appraisal	**Unit of assessment** **Performance standards** **Analysis and interpretation of performance data** Weighing the domains Weighing the indicators Calculations Weighing the quality scores Differentiation of quality scores Feedback
Reimbursement	**Financial rewards** Payment Size of the bonus Spending the bonus

Effect of P4P programs

The effectiveness of P4P programs is still inconclusive[20,21], despite the proliferation of these programs. Petersen et al. (2006) give some suggestions for possible elements of successful P4P.[20] They suggest to use combined incentives for both overall improvement and achievement of a threshold and to make use of process and outcome indicators as target measures. Most studies reflect on the effects on specific indicators, but seldom reflect on the psychology of how people respond to incentives. Mehrotra et al. (2010) presented seven lessons from behavioural economics that might enhance the effectiveness of P4P programs[22]:

1 A series of small incentives is better than one large incentive
2 A series of tiered absolute thresholds is better than one absolute threshold
3 Reducing the lag times between care and receipt of incentives increases the behavioural response
4 Although withholds have more of an effect than bonuses, one needs to be cognizant of the negative psychological response
5 Reducing the complexity of an incentive plan increases the behavioural response

6 P4P program and incentive payments should be decoupled from usual reimbursement
7 'In kind' rewards may be a stronger drive of change than a cash reward of the same

A systematic use of behavioural economics as described is lacking in the current P4P programs. More research into elements of effective P4P is needed.

Experiences with P4P programs

Studies about the experiences of target users with P4P programs are scarce. Experiences of target users that were involved in designing the P4P program are unknown, though involving target users could enlarge P4P's effectiveness.[15] Physicians participating in P4P programs in Massachusetts and California showed positive attitudes toward P4P, but were ambivalent about specific features of these programs.[23] General practices participating in the Australian Practice Incentives Program (PIP) were asked about their views on PIP's contribution to quality of care and improved access. Their views were mixed, with 27 percent of providers responding that PIP gives significant benefit to their practice, 36 percent responding that there is medium benefit, and 27 percent responding that the benefit is minor.[24] Campbell et al. (2008) interviewed GPs and nurses about their views on changes in healthcare as a result of the Quality and Outcomes Framework (QOF).[25] The respondents believed that the financial incentives had been sufficient to change behaviour and to achieve targets, but they also mentioned some unintended consequences such as a decline in personal continuity of care. Furthermore, the interviewees worried about an ongoing culture of performance monitoring in the UK. So, further research into experiences of P4P target groups is demanded.

A participatory P4P program

The assessment of different P4P programs so far, show the importance of involvement of target users during the complete phases of the development, implementation, and evaluation of P4P.[15] In four studies that involved target users, three different P4P programs were evaluated and improvements of 20% on average in three to five years were reported.[26-29]

Another important message is that the design of the P4P program should be developed along three framework components: performance measurement, appraisal and reimbursement.[8] In defining the appraisal the seven lessons from behavioural economics that might enhance the effectiveness of P4P programs should be taken into account.[22]

At the beginning of this chapter it was explained that improving the quality of care was instigated by stimulating the intrinsic motivation. Later on extrinsic interventions seem to took over. It can be questioned whether it is possible to

develop a P4P program that focuses on both intrinsic and extrinsic motivation. This would result in healthcare providers working on quality of care based on professionalism, and being rewarded with a financial incentive for their performance on quality of care.
In this thesis we attempted to develop a structured participatory P4P program with both intrinsic and extrinsic components. The mainly intrinsic component will be realized by linking the P4P program to the Dutch National Accreditation program[30]; and the performance payment relates to the extrinsic component mainly.

Objective and study aims

The aim of this study was to develop a P4P program with active involvement of the target users aimed at improving quality of care and rewarding professionals for their performance. This thesis will address the process of developing a P4P program, discussing the design of performance measurement, appraisal and reimbursement. We will also cover the validity and reliability of the clinical care performance measures. The participatory P4P program was evaluated on its effect and on the experiences of the target users. Since quality improvement is the main issue, and the rationale in P4P programs is that feedback on actual data and improvement plans can improve quality of care, a separate study was designed in which the impact of feedback on actual performance in general practice was studied.

The research questions of the separate studies were as follows:

1 What is the impact of feedback on performance accompanied by information on best practices on the accessibility and availability in general practice? (chapter 2)
2 Which design choices are made by the target users for a P4P program, in which the different options for performance measurement, appraisal and reimbursement were discussed in a systematic consensus procedure? (chapter 3)
3 What is the validity of the clinical care domain? (chapter 4)
4 What is the effect of the P4P program on clinical indicators as well as on the patient experience after one year? (chapter 5)
5 Has the involvement of target users in designing the P4P program led to positive evaluations and confidence in its future use? (chapter 6)

References

1
Grol R, Wensing MJP, Eccles M. *Improving patient care: The implementation of change in clinical practice*, 2005.

2
Kunzi B, Borge B, Van den Hombergh P. The role of feedback for quality improvement in primary care. *Quality Management in Primary Care*, 2007.

3
Hayes RP, Ballard DJ. Review: feedback about practice patterns for measurable improvements in quality of care–a challenge for PROs under the Health Care Quality Improvement Program. *Clin Perform Qua. Health Care* 1995;3(1):15-22.

4
Jencks SF. Changing health care practices in Medicare's Health Care Quality Improvement Program. *Jt Comm J Qual Improv* 1995;21(7):343-47.

5
Jencks SF, Wilensky GR. The health care quality improvement initiative. A new approach to quality assurance in Medicare. *JAMA* 1992;268(7):900-03.

6
Kiefe CI, Allison JJ, Williams OD, Person SD, Weaver MT, Weissman NW. Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. *JAMA* 2001;285(22):2871-79.

7
Mandel KE. Aligning rewards with large-scale improvement. *JAMA* 2010;303(7):663-4.

8
Mannion R, Davies HT. Payment for performance in health care. *BMJ* 2008;336(7639):306-08.

9
Rowe JW. Pay-for-performance and accountability: related themes in improving health care. *Annals of internal medicine* 2006;145(9):695-9.

10
Roland M. Linking physicians' pay to the quality of care--a major experiment in the United kingdom. *N Engl J Med* 2004;351(14):1448-54.

11
Association IH. 2013. Integrated Healthcare Association, 2013, http://iha.org/p4p_california.html, year cited 2013, date cited 15 February

12
Australia M. 2013. Medicare Australia, 2013, http://www.medicareaustralia.gov.au/provider/incentives/pip/index.jsp, year cited 2013, date cited 15 February

13
Maffei RM, Turner N, Dunn K. Building blocks to adopting a pay-for-performance model. *JONAS Healthc Law Ethics Regul* 2008;10 (3):64-69.

14
Rosenthal MB, Fernandopulle R, Song HR, Landon B. Paying for quality: providers' incentives for quality improvement. *Health Aff(Millwood)* 2004;23(2):127-41.

15
Van Herck P, De SD, Annemans L, Remmen R, Rosenthal MB, Sermeus W. Systematic review: Effects, design choices, and context of pay-for-performance in health care. *BMC Health Serv Res* 2010;10:247.

16
Mannion R, Davies H, Marshall M. Impact of star performance ratings in English acute hospital trusts. *J Health Serv Res Policy* 2005; 10(1):18-24.

17
Gruen RL, Elliott JH, Nolan ML, Lawton PD, Parkhill A, McLaren CJ, et al. Sustainability science: an integrated approach for health-programme planning. *Lancet* 2008;372(9649):1579-89.

18
Scott IA. Pay for performance in health care: strategic issues for Australian experiments. *Med J Aust* 2007;187(1):31-35.

19
Campbell SM, Braspenning J, Hutchinson A, Marshall MN. Research methods used in developing and applying quality indicators in primary care. *BMJ* 2003;326(7393):816-19.

20
Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S. Does pay-for-performance improve the quality of health care? *Ann Intern Med* 2006;145(4):265-72.

21
Rosenthal MB, Frank RG. What is the empirical basis for paying for quality in health care? *Med Care Res Rev* 2006;63(2):135-57.

22
Mehrotra A, Sorbero ME, Damberg CL. Using the lessons of behavioral economics to design more effective pay-for-performance programs. *Am J Manag Care* 2010;16(7):497-503.

23
Young GJ, Meterko M, White B, Bokhour BG, Sautter KM, Berlowitz D, et al. Physician attitudes toward pay-for-quality programs: perspectives from the front line. *Med Care Res Rev* 2007; 64(3):331-43.

24
Office ANA. *Practice Incentives Program*. Department of Health and Ageing. Medicare Australia.

Audit Report No.5 2010-11, 2010.

25
Campbell SM, McDonald R, Lester H. The experience of pay for performance in English family practice: a qualitative study. *Ann Fam Med* 2008;6(3):228-34.

26
Amundson G, Solberg LI, Reed M, Martini EM, Carlson R. Paying for quality improvement: compliance with tobacco cessation guidelines. *Jt Comm J Qual Saf* 2003;29(2):59-65.

27
Chung RS, Chernicoff HO, Nakao KA, Nickel RC, Legorreta AP. A quality-driven physician compensation model: four-year follow-up study. *J Healthc Qual* 2003;25(6):31-37.

28
Gilmore AS, Zhao Y, Kang N, Ryskina KL, Legorreta AP, Taira DA, et al. Patient outcomes and evidence-based medicine in a preferred provider organization setting: a six-year evaluation of a physician pay-for-performance program. *Health Serv Res* 2007;42 (6 Pt 1):2140-59.

29
Larsen DL, Cannon W, Towner S. Longitudinal assessment of a diabetes care management system in an integrated health network. *J Manag Care Pharm* 2003;9(6):552-58.

30
V
an den Hombergh P, Schalk-Soekar S, Kramer A, Bottema B, Campbell S, Braspenning J. Are family practice trainers and their host practices any better? comparing practice trainers and non-trainers and their practices. *BMC family practice* 2013;14:23.

Chapter 2

Improving access to primary care: the impact of a quality-improvement strategy

Kirsten Kirschner
Jozé Braspenning
Irma Maassen
Angelique Bonte
Jako Burgers
Richard Grol

Quality and Safety in Health Care 2010;19(3):248-251.

Abstract

Problem Many patients are not satisfied with the accessibility and availability of general practice, and they would like to see improvement.

Design Quality-improvement study with pre-intervention and post-intervention data collection in 36 general practices.

Setting General practices located in the south of the Netherlands.

Key measures for improvement Patient satisfaction, experiences and awareness; practice information; and experiences of a mystery patient.

Strategy for change The practices received feedback about their accessibility and availability compared with data from practices of colleagues. The practices developed practice-based improvement plans using these feedback results.

Effects of change Eighty per cent of the improvement plans were completed or almost completed in 5 months. After the intervention, the accessibility by phone within 2 min increased significantly (10% improvement). The practices that designed an improvement plan showed a larger increase (25% improvement) than practices that did not. Patient awareness of an information leaflet and a separate telephone number for emergency calls also significantly increased (29% improvement and 12% improvement) in practices that designed improvement plans.

Lessons learned Feedback and practice-based improvement plans were a stimulus to work on and to improve accessibility and availability. All practices started improvement plans, but the overall effect of the changes was modest. This may be due to acceptable accessibility and availability before the intervention was introduced and to the time period of 5 months, which seemed to be too short to complete all practice-based improvement plans. The mystery patient was more satisfied with the accessibility than the real patients. This may be related to our concept of accessibility. We learned that adding a mystery patient for data collection can contribute to more objective measurements of practice accessibility than patient questionnaires alone.

Access to healthcare is a prerequisite for quality of primary care services. The concept of 'access' includes availability, accessibility, accommodation, affordability and acceptability.[1] In this study, we focus on accessibility and availability in general practice during practice hours for both routine care and emergency care.[2] Waiting time on the phone for ordinary consultations is one of the aspects of accessibility. Making an appointment within two working days with your general practitioner is one of the aspects of availability. Patients appreciate a doctor who is available within a short time. In particular, in primary care, fast access may contribute to the perception of patient-centred healthcare. Access to primary healthcare services is a public and political concern in several countries.[3] The National Health Service plan in the United Kingdom, for instance, includes access to primary care as one of the key components.[4] The General Medical Services contract included performance indicators for access in 2006-2007.[5] Research in general practice in the Netherlands shows that patients were satisfied with the accessibility in general but less satisfied with certain aspects such as waiting time for an appointment, accessibility by phone, being able to speak to the practitioner on the telephone, waiting time in the waiting room, and emergency care services.[6-8] A strategy to improve the access to general practice consists of auditing of and feedback about actual services. Feedback about individual performance compared with that of peer performance can be a powerful motivator for change.[9-12] The aim of auditing and feedback is to present data about practice performance and to encourage practices to design specific practice-based improvement plans as a guide for change.[13] In addition to the performance data of individual practices, information about best practices was provided.[14] In this study, we examined the impact of this approach on the accessibility and availability in general practice.

Key measures for improvement

Three measures were used for the evaluation of the intervention:

1 Patient satisfaction, experience and awareness
2 Information about general practice services
3 Experiences of a mystery patient.

Process of gathering information

Design and participants

We asked 129 general practices in the south of the Netherlands to participate in our project. Sixty-six practices responded of which 36 agreed to participate voluntarily (61%). In each practice, 40 patients registered with the general practice were asked to participate. These patients were randomly selected. We collected data about accessibility and availability using: (1) questionnaires completed by adult patients (>18 years), (2) questionnaires about general practice

services completed by general practitioners and (3) feedback from one mystery patient who made 15 calls for ordinary consultations and 5 emergency calls. When the phone was busy or not answered, the phone numbers were called three more times. When the phone was picked up and the patient was asked to hold, the service was classed as accessible. Data were collected before the intervention and 5 months after the intervention.

Outcome measures

The measures used were waiting time on the phone for emergencies and ordinary consultations, waiting time in the waiting room before consultation, waiting time for an appointment (both acute and chronic illnesses) and the quality of information service. Patients were asked to report their satisfaction and experience with the phone accessibility for emergencies and ordinary consultations, waiting time for an appointment and the waiting time in the waiting room before consultation. A specific part of an internationally validated questionnaire covering access of care (Europep, Visitation Instrument Practice Management) was used to collect the data.[15,16] Practice information was measured using the same instrument.[16] The patients' awareness of an information leaflet, a telephone number for emergencies and information about waiting time in the waiting room were also asked for. The general practitioner was asked to provide practice information concerning the presence of an information leaflet, a practice website and a specific telephone number for emergencies.

The mystery patient was asked to make an appointment for an ordinary consultation and another for an emergency. Good phone accessibility for an ordinary consultation was defined as receiving personal contact within 60 s or when the answering machine gave an alternative telephone number within 60 s. For emergencies, good accessibility was defined as receiving personal contact within 30 s or when the answering machine gave an alternative telephone number within 30 s. When no contact had been established after three attempts, the service was classed as inaccessible at that time.

Data collection

The patients completed the questionnaires after their consultations.

A practice assistant handed out the questionnaires during one week. Patients that filled in a questionnaire could deposit this questionnaire in a closed box. After one week, all questionnaires that were filled in were sent to the research team. One general practitioner in each practice filled in the questionnaire about practice services. The mystery patient called for an ordinary consultation three times a day for 5 days and for emergency services called once a day for 5 days. The measurement of patient satisfaction, experience and awareness; practice information; and the mystery patient's investigation were repeated after 5 months

during which the practice could have improved their performance using the individual feedback and information on best practices. At the end of the study, the process was evaluated by means of a questionnaire to investigate the experiences of the practices with this study. The questions covered the distribution of the patient questionnaires, the mystery patient's investigation, the quality of the feedback, and the information on best practices.

Strategy for change: feedback and encouraging design of pratice-based improvement plans

The practices received individual feedback from the research team based on their accessibility and they received information about three best practices. These best practices were selected based on the most positive patient satisfaction with waiting times in the waiting room, accessibility by phone, and information service. The recommendations for practice-based improvement plans were suggested to the general practitioner and concerned all practice members. Recommendations were made when practice performance was 5% lower than the mean of all participating practices or when the individual performance was relatively low (patient satisfaction, experience or awareness <75% for one specific subject). Dutch literature shows that the average perceived accessibility of patients is 80%.[6] We accepted a deviation of 5%.
Two and 5 months after designing the improvement plans, the practices were called to report on their progress.

Analysis and interpretation

We grouped the improvement plans into four categories: completed, ongoing, not started and cancelled. The focus of the improvement plans included:

1 Information service
2 Accessibility by phone
3 Waiting time in the waiting room
4 Organisation of consulting hours
5 Consulting hours by phone.

We examined the overall impact of the quality-improvement intervention on patient satisfaction and experience, including those of the mystery patient. An independent sample t test was used to analyse the data regarding the impact of the quality improvement strategy. Only quality-improvement plans implemented in more than two practices were selected for this analysis.

Study population

At baseline, 1256 patients from the 36 general practices filled in the questionnaire. The response rate was 87.2%. Only one of the 36 practices originally

recruited dropped out because we did not succeed to make an appointment with the general practice for discussion about their performance during the study period. After 5 months, 1071 (76.5%) patients from 35 practices filled in the patient questionnaire. The practices included in the study were representative for all Dutch general practices (table 1).[17] There was a small over-representation of group practices and practices in rural areas in comparison with the overall Dutch general practices.

Patients aged 44-64 years were over-represented in the study population. Women in the study population were over-represented in comparison with Dutch general practices.

Table 1 Practice characteristics of the study population in comparison with all Dutch general practices

	Study population 36 practices		*All Dutch general practices*[1] *4455 practices*	
	Number	**%**	**Number**	**%**
Practice type				
Solo	13	37.1		47.9
Dual	12	34.3		30.1
Group/health centre	10	28.6		22.0
Urban area				
Large city	13	36.2		45.9
Small city	18	50.0		41.8
Rural city	5	13.9		12.3
Practice size				
– Mean patient population per practice	4349		4283[2]	
– Mean patient population per full-time equivalent general practitioner	2475		2437[2]	

[1] Since 1/1/2006. Source: NIVEL(17)
[2] Sourche: NIVEL, NS2 (65 practices) (18)
Large city: >1500 adresses per km²; small city: 500-1500 addresses per km²; rural area: <500 addresses per km²

Effects of change

The 36 participating practices developed 123 practice-based improvement plans using the practice feedback results and practice information. The practices perceived the feedback about their performance as very useful. After 5 months, 26 practices filled in a questionnaire concerning the progress of their improvement plans. Table 2 shows that almost half (53/123) of the improvement plans were related to information service. One fourth (31/123) was related to phone accessibility, and another one fourth (33/123) to waiting time in the waiting room. Six plans were related to the consulting hours. After 5 months, 51 (42%) improvement plans were completed.

Table 2 Progress of practice-based improvement plans

Module	*Completed* N[1]	*Ongoing* N[1]	*Not started* N[1]	*Cancelled* N[1]	*Total plans* N[1]
Information service	22	20	11	0	53
Accessibility by phone	11	14	5	1	31
Waiting time in the waiting room	15	16	1	1	33
Organisation consulting hours	0	0	2	0	2
Consulting hours by phone	3	0	1	0	4
Total	**51**	**50**	**20**	**2**	**123**

[1] N, number of practice-based improvement plans
There were 26 practices

Table 3 presents the impact of four specific practice-based improvement plans. The plans 'distribution of information leaflet', 'publicity of the telephone number for emergencies' and 'phone accessibility for an ordinary consultation within 2 min' improved significantly.

Table 3 Impact of practice-based improvement plans

Module	*Practice-based improvement plan*	*Number of practices*	*Indicator*	*To*		*T1*		*Change*
				%	N[1]	%	N[1]	%
Information service	Distribution of information leaflet	5	Patients' awareness	58.2	170	87.3	150	29.1*
	Publicity of telephone number for emergencies	7	Patients' awareness	66.5	233	78.8	189	12.3*
Accessibility by phone	Accessibility by phone for usual consultation	3	Good/excellent	37.6	101	29.0	107	- 8.6
			Contact within 2 minutes	35.4	96	60.2	103	24.8*
Waiting time in the waiting room	Delay in consulting hours	12	Delay information according to patients	29.1	261	21.2	208	- 7.9

* Significance: $p < 0.05$
[1] N, Number of patients that filled in a questionnaire and filled in this question
To Pre-intervention
T1 Post-intervention

Table 4 shows that there were significant changes in patient experience of phone accessibility for ordinary consultation and for making an appointment within two working days. There was also a significant change in patient satisfaction with the waiting time in the waiting room.

Table 4 Impact of interventions

			To		*T1*		*Change*
			%	N[1]	%	N[1]	%
Accessibility by phone (emergency)	Patient satisfaction	Good / excellent	57.6	533	55.7	494	-1.9
	Patient experience	Contact within 2 min	82.9	251	80.9	173	-1.9
	Mystery patient	Contact within 30 s	94.0	-	94.0	-	-
Accessibility by phone (ordinary consultation)	Patient satisfaction	Good / excellent	48.6	1191	46.5	1032	-2.1
	Patient experience	Contact within 2 min	61.0	1164	71.3	963	10.3*
	Mystery patient	Contact within 1 min	84.0	-	87.0	-	3.0
Waiting time in the waiting room	Patient satisfaction	Good / excellent	42.2	1156	48.3	988	6.1*
	Patient experience	Within 10 min	64.1	1185	65.4	1007	1.3
Consulting hours	Patient satisfaction	Good / excellent	56.6	1175	59.4	1005	2.8
	Patient experience	Within two working days	94.0	1184	91.3	998	-2.7*

* Significance: $p < 0.05$
[1] N, Number of patients that filled in a questionnaire and filled in this question
To Pre-intervention
T1 Post-intervention

Lessons learned

All participating general practices were motivated to improve their accessibility and availability before the study started. They found the feedback about their performance very useful. Most of the practice-based improvement plans concerned information service, waiting time in the waiting room and phone accessibility. After 5 months, 80% of the plans were completed or almost completed. The improvement plans concerning process improvements would likely be completed after a longer period. The type of improvements concerned structural changes in daily practice and routines, which are expected to be sustainable.

The experience of the mystery patient and patient satisfaction showed that the accessibility in general practice was good. These results are consistent with other studies that show high scores on patient satisfaction in the Netherlands.[7,8] It can be argued that our results are an over-estimation as we defined accessibility as 'good' when an alternative phone number was provided and did not restrict accessibility to speaking to a person since we assumed that a patient knows how to handle according to this information. Over-estimation of patient satisfaction could also be due to the patient population, which included relatively more patients with chronic disease. These patients visit the practice more often and know the preferred times to call the practice and may, therefore, be more satisfied than other patients.

Adding a mystery patient for data collection could contribute to more objective measurements of practice accessibility than patient questionnaires alone. In our study, the mystery patient was more satisfied with the accessibility than the real patients. It may be argued that we only included one mystery patient in our

study. We gave clear instructions to this 'patient', which could explain more patient satisfaction. Another explanation is related to our concept of accessibility, which was defined as acceptable if the mystery patient contacted the practice after three attempts at most. In contrast, patients might expect that there should be personal contact directly after the first call. If more information is available on patient expectations, specific interventions could improve satisfaction with services subsequently.[18] Finally, practices might have anticipated on a call from a mystery patient. However, they had no idea when the mystery patient would call during the study period, so anticipation is unlikely.

A substantial proportion (43%) of the improvement plans dealt with practice information services. This can be very effective in enhancing a patient's knowledge on when and how the general practice is accessible. Increasingly, practices use websites for information services, which can be easily updated with the latest information on access and availability. Future research on accessibility could address the use of websites and other alternative information services. Future research should also take into account that there is a gap between the perceptions of a mystery patient and an actual patient concerning accessibility.

References

1
Penchansky R, Thomas JW. The concept of access: definition and relationship to consumer satisfaction. *Med Care* 1981;19:127e40.

2
Sips SBI, Tielens VCL, Van der Voort JPM. Bereikbaarheid/beschikbaarheid: NHG-Standaard [Accessibility and availability: the Dutch College of General Practitioners guideline]. *Huisarts Wet* 1989;32:219-22.

3
Rubin G, Bate A, George A, et al. Preferences for access to the GP: a discrete choice experiment. *Br J Gen Pract* 2006;56:743-8.

4
Department of Health. Delivering the NHS plan. Norwich: The Stationery Office, 2002.

5
Focus on access (England) 2006/07. British Medical Association, 2007. http://www.bma.org.uk/images/ focus%20on%20access%202006-07_tcm41-37220.pdf

6
Engels Y. *Assessing and improving management in primary care practices in the Netherlands and in Europe*. Nijmegen: Radboud University Nijmegen, 2005.

7
Grol R, Wensing M, Mainz J, et al. Patients in Europe evaluate general practice care: an international comparison. *Br J Gen Pract* 2000;50:882-7.

8
Wensing M, Vedsted P, Kersnik J, et al. Patient satisfaction with availability of general practice: an international comparison. *Int J Qual Health Care* 2002;14:111-8.

9
Hayes RP, Ballard DJ. Review: feedback about practice patterns for measurable improvements in quality of careda challenge for PROs under the Health Care Quality Improvement Program. *Clinical Performance and Quality Health Care* 1995;3:15-22.

10
Jencks SF, Wilensky GR. The health care quality improvement initiative. A new approach to quality assurance in Medicare. *JAMA* 1992;268:900-3.

11
Jencks SF. Changing health care practices in Medicare's Health Care Quality Improvement Program. *The Joint Commission Journal on Quality Improvement* 1995;21:343-7.

12
Kiefe CI, Allison JJ, Williams OD, et al. Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. *JAMA* 2001;285:2871-9.

13
Kunzi B, Borge B, Van den Hombergh P. The role of feedback for quality improvement in primary care. In: Grog R, Dautzenberg M, Brinkmann H, eds. *Quality Management in Primary Care; European Practice Assessment*. Bielefeld, Germany: Verlag Bertelsmann Stiftung 2004:106e28.

14
Edgman-Levitan S, Dale Shaller PA, McInnes K, et al. *The CAHPS improvement guide practical strategies for improving the patient care experience*. Boston: Department of Health Care Policy Harvard Medical School, 2003.

15
Wensing MJP. *Patients evaluate general practice*. Nijmegen: Catholic University of Nijmegen, 1997.

16
Van den Hombergh P. *Practice visits. Assessing and improving management in general practice*. Nijmegen: Catholic University of Nijmegen, 1998.

17
Nivel-beroepenorganisaties. 2007. www.nivel.nl.

18
Saxton JW, Finkelstein MM. Expectation management to reduce liability risk. 2005. http://www.physiciannews.com/business/905saxton.html

19
Schellevis FG, Westert GP, De Bakker DH, et al. Tweede nationale studie naar ziekten en verrichtingen in de huisartsenpraktijk [Second national study of diseases and performance in general practice]. Utrecht/Bilthoven: Nivel/RIVM, 2004.

Chapter 3

Design choices made by target users for a pay-for-performance program in primary care: an action research approach

Kirsten Kirschner
Jozé Braspenning
JE Annelies Jacobs
Richard Grol

BMC Family Practice 2012, 13:25

Abstract

Background International interest in pay-for-performance (P4P) initiatives to improve quality of health care is growing. Current programs vary in the methods of performance measurement, appraisal and reimbursement. One may assume that involvement of health care professionals in the goal setting and methods of quality measurement and subsequent payment schemes may enhance their commitment to and motivation for P4P programs and therefore the impact of these programs. We developed a P4P program in which the target users were involved in decisions about the P4P methods.

Methods For the development of the P4P program a framework was used which distinguished three main components: performance measurement, appraisal and reimbursement. Based on this framework design choices were discussed in two panels of target users using an adapted Delphi procedure. The target users were 65 general practices and two health insurance companies in the south of the Netherlands.

Results Performance measurement was linked to the Dutch accreditation program based on three domains (clinical care, practice management and patient experience). The general practice was chosen as unit of assessment. Relative standards were set at the 25th percentile of group performance. The incentive for clinical care was set twice as high as the one for practice management and patient experience. Quality scores were to be calculated separately for all three domains, and for both the quality level and the improvement of performance. The incentive for quality level was set thrice as high as the one for the improvement of performance. For reimbursement, quality scores were divided into seven levels. A practice with a quality score in the lowest group was not supposed to receive a bonus. The additional payment grew proportionally for each extra group. The bonus aimed at was on average 5% to 10% of the practice income.

Conclusions Designing a P4P program for primary care with involvement of the target users gave us an insight into their motives, which can help others who need to discuss similar programs. The resulting program is in line with target users' views and assessments of relevance and applicability. This may enhance their commitment to the program as was indicated by the growing number of voluntary participants after a successfully performed field test during the procedure. The elements of our framework can be very helpful for others who are developing or evaluating a P4P program.

Background

International interest in pay-for-performance (P4P) initiatives to improve quality of health care is growing. Despite the proliferation of P4P programs, the evidence to support their use is still inconclusive.[1,2] One of the reasons may be the differences between P4P programs. Incentives in current programs vary in terms of number and type of indicators, professional standards and quality domains (clinical care, patient experience, organisation of care).[3-7] The size of the incentive and the unit of assessment in P4P programs can influence their effectiveness.[8] Experiences with different P4P programs led to a framework for design choices regarding the P4P approach. Three essential framework components to design a P4P program can be distinguished: performance measurement, appraisal and reimbursement.[9-12] The performance measurement should include valid and reliable indicators that make sense to the target group. Appraisal in a P4P program means defining the unit of assessment and the performance standards, but also describing the analysis and interpretation of the data. Based on the analysis and interpretation of the data a reimbursement system can be built.[10] Another remarkable feature of current P4P programs is that they are mostly designed and implemented top-down by policy makers and managers.[13] P4P programs can be seen as an innovation in care, and it is known that the sustainability of an innovation can be improved by involving target users.[14] It has also been suggested to involve target users in the developmental process of a P4P program, because this can contribute to the effect of incentivised indicators.[15,16] A more bottom-up procedure in designing a P4P program may improve its future implementation and its effectiveness.

Evaluation of the involvement of target users in the decisions about the P4P program may contribute to the growing field of P4P research. One may assume that involvement of health care professionals in the goal setting and methods of quality measurement and subsequent payment schemes may enhance their commitment to and motivation for P4P programs and therefore the impact of these programs. Nevertheless, you have to reckon with conflicts of interest when involving target users. Therefore it is important to develop the P4P program in a systematic way, such as the Delphi procedure.[17] The perspectives of all target users become distinct, and the decisions made are transparent for the target users. The aim of our study was to design a P4P program using a bottom-up procedure, in which the different options for performance measurement, appraisal and reimbursement were discussed by the target users in a systematic consensus procedure. We will present this bottom-up process of development of the P4P program and its resulting design.

Methods

The design options in the P4P framework

We searched the literature for relevant elements for our P4P program, to be discussed by the target users. Table 1 gives an overview of the elements and design options.

The *performance indicators* covered three domains, clinical care, practice management and patient experience, and were derived from the Dutch National Accreditation Program for general practices.[18] The target users were asked whether these three domains, the subjects and the indicators were appropriate for the P4P program. For clinical care the target users could choose from indicators for diabetes, COPD, asthma, cardiovascular risk management, influenza vaccination, cervical cancer screening and prescribing acid suppressive drugs and antibiotics. For practice management, which is measured with the validated Visitation Instrument Practice management (VIP)[19], they could approve various indicators for infrastructure, team, information, and quality and safety. The indicators for patient experience to agree on were based on the internationally validated EUROPEP instrument[20], which evaluates both the general practitioner and the organisation of care. Furthermore, target users could decide on collecting data for all three domains each year versus a trimmed-down version of the program.

The *appraisal* and *reimbursement* elements and options to be discussed were derived from the literature.[1,8,10,11] The following design elements of P4P programs were described: unit of assessment, performance standards, analysis and interpretation of performance data, and financial rewards. The options for the unit of assessment, either the general practitioner, the general practice or a larger organisational unit, were presented together with evidence that the smaller the unit the stronger the stimulation of quality improvement[1], and the practical consideration that the general practice is the unit of assessment within the Dutch National Accreditation program. The options presented for performance standards were either absolute or relative performance standards.[10] Most existing programs are based on absolute standards.[1,8,21] The target users were asked whether performance standards should vary between indicators/subjects. Some indicators might need lower minimum standards because they are more difficult to reach than others. Concerning the analysis and interpretation of performance data the options were to weigh domains and indicators either differently or to weigh them equally. In the Quality and Outcomes Framework (QOF), for instance, performance on clinical indicators receives more weight than practice management or patient experience.[7] For calculating quality scores options were to either calculate a quality score for each domain separately or to calculate one overall domain-score. Moreover the target users could choose whether both the quality level and the improvement of performance should

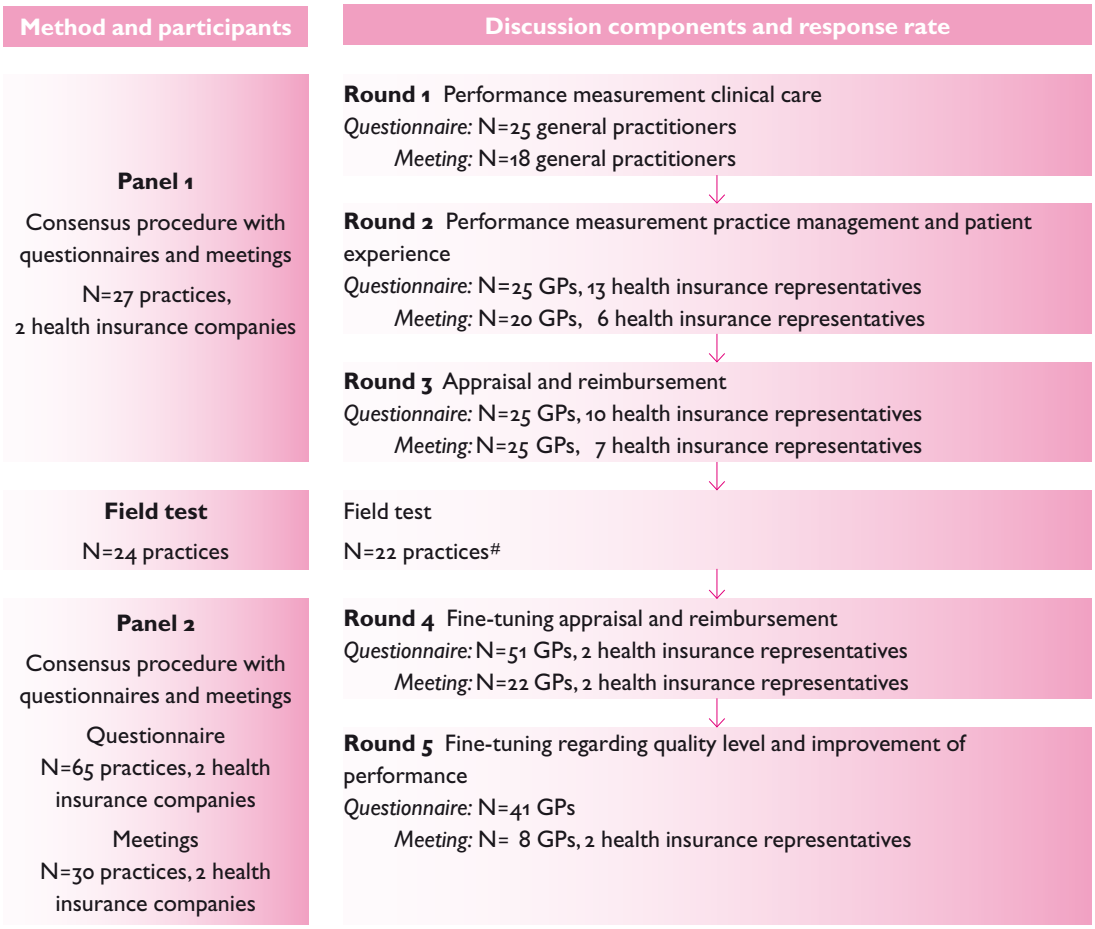
be incentivised and whether to weigh the scores differently or equally. A combination of incentives for both the quality level and improvement of performance will encourage both low and high performing providers to improve quality.[1,16] In order to link a bonus to the quality, quality scores need to be differentiated into levels. The options given were: 4 levels (quartiles), 5 to 7 levels, or 8 to 10 levels. The more levels, the more smaller improvements will be worth the investment. For the feedback a discussion was started on a proper benchmark and on risk adjustment. The options presented for a benchmark were the median, the best practice (75th percentile or 90th percentile) or a combination. Improvement can best be stimulated by feedback in a reachable range[22], thus practices with relatively low scores are stimulated by the average of the peer group, and practices with high scores by best practices. Comparing practices with others without appropriate risk adjustment can be misleading. Risk factors include patient demographic and/or clinical factors, which can influence outcomes of care. The target users had to decide on risk adjustment of the indicator scores, which is either to adjust the benchmark (indirect correction) or the indicator scores (direct correction). Concerning the feedback the target users could choose either a one-step procedure or a two-step procedure. In the one-step procedure practices receive feedback and bonus together. In the two-step procedure practices first receive feedback, and receive the bonus only after they have had the opportunity to respond to their feedback. Concerning the *reimbursement* the options for the method of payment were either money, human resources, a sabbatical leave or a combination of these. For the size of the bonus we asked the target users whether an average bonus of 5000 to 10000 Euros (depending on practice size), which is on average 5% to 10% of the practice income, would be appropriate. In other P4P programs the size of the bonus varied between US$ 2 per patient and US$ 10000 per practice.[23] The size of the bonus should not be too small as this may limit the effects, but neither should it be too high because of unintended consequences like gaming.[21,24,25] The options for spending the bonus were either without obligations or with obligations (spending for the practice related to a goal, possibly preset) or a combination of these options.

Study design

An action research approach[26] was applied with participation of future target users in the development of the P4P program. To reach consensus an adapted Delphi procedure[27] was used in two panels of target users. (Figure 1) The target users in question were general practitioners (GPs) and payers (representatives of health insurance companies). General practices in the south of the Netherlands were invited by the two regional health insurance companies to participate voluntarily in this P4P experiment. We aimed at participation of 20 to 25 general

practices, and at least one representative of each health insurance company. To achieve consensus on the P4P design, two rounds were organized to discuss the methods of performance measurements (one on clinical care, and one on practice management and patient experience) and one round to discuss the methods of appraisal and reimbursement. The participating practices were also invited to volunteer in a field test in which data were collected based on the previous choices for the measurement of clinical performance, practice management and patient experiences. Feedback to the practices was delivered and the resulting bonus was paid according to the system agreed on. After the field test the panel was extended with general practices that were also willing to participate in this P4P experiment. In this second panel we discussed the methods of appraisal and reimbursement based on the results of the field test (round four) and the design options regarding quality level and improvement of performance (round five) to fine-tune the P4P program.

Figure 1 Procedure design selection of a P4P program by target users



Two practices dropped out, one due to illness and the other due to disassociation of practice owners

Consensus procedure

In each round a written questionnaire with the design options for the P4P program was sent to the target users two weeks before the planned meeting. In the questionnaire they were provided with background evidence on the options as described in the section 'The design options in the P4P framework' and they were asked to make a choice. Each meeting started with explaining the aim of the discussion and feedback on the results of the questionnaires. All design options were discussed, but for the performance indicators the project team decided not to discuss indicators with high consensus, defined as less than 30% or more than 70% in favour. At the end of each meeting the panel members completed the same questionnaire again. The decision rule for inclusion of clinical indicators was set at more than 70% in favour, and for the other design options a majority rule was applied.

All panel meetings were held in the region in question to enhance participation. Payers and GPs attended the same discussion meetings which lasted 2 hours. The general practices in the first panel received 1500 Euros for participating in the panel as well as in the field test. Each GP in the second panel received 100 Euros for attending the meetings.

Results

Study population

The number of general practitioners and health insurance representatives that filled in the questionnaires and attended the meetings for the specific panels are presented in Figure 1. The number of GPs that could attend the meetings in round four and five were restricted to 30 due to the large number of practices that voluntarily participated in the P4P program.

In panel 1 the response rate for the GPs was on average 93% for the questionnaires and 78% for the meetings, and in panel 2 71% and 50% respectively. The health insurance representatives decided to leave the discussion on the performance indicators to the experts (GPs). They participated amply in panel 1 and their participation decreased in panel 2.

Design choices

The successive panel procedures and the field test resulted in a P4P program which is presented in Table 1.

Performance measurement

The target users thought clinical care, practice management and patient experience to be appropriate domains for the P4P program, as well as the subjects within these domains (see Table 1). Some GPs remarked that the *clinical conditions* to be assessed were mainly focused on chronic care, though GP care comprises

much more. Especially communication skills were missed. Although patient experiences were to be assessed, some GPs stated that communication was not reflected enough in the indicators. The GPs also discussed the fact that choosing indicators would result in a certain focus that could distract them from the more general goal of quality improvement. Practices may concentrate their performance on the indicators from the P4P program. It was proposed that in the long-term a large set of indicators will be needed. Then the P4P program could have different sets for different years. Some GPs even suggested that the practice should not be aware of the existing set. For clinical care, GPs were convinced that the outcomes were a mixed result of patient and doctor's performance. It was therefore decided that the payment should be based on the process measures only, but the GPs would like to receive feedback on the outcome indicators as well. So, data for both process and outcome indicators were collected.

Although the health insurance representatives stated they would leave the decisions on clinical care options to the GPs, they joined the discussion in the meetings on the prescription indicators. The prescription indicators were highly valued by the health insurance representatives. The GPs questioned these indicators which resulted in not including the acid suppressive drugs indicators in the program, and including indicators on prescribing antibiotics. Some items of *practice management* were excluded due to their estimated low correlation with the quality of practice management such as financial accounting. The EUROPEP instrument, consisting of 23 items measuring *patient experience*, was supplemented with four items regarding the possibility to ask for a longer consultation, accessibility by phone, getting an appointment with your own doctor and an accessible procedure for complaints. All selected quality indicators can be found in the additional file.

The target users agreed that at *baseline*, data should be collected for all three domains. For the *years to follow* the data collection for the practice management domain was judged to be unnecessary as it is not likely that this will change substantially over two or three years and as the data collection in this domain is very labour intensive.

Appraisal

The general practice was chosen as *unit of assessment*, because in the context of the P4P program the incentive would be targeted at this level. Since the data on clinical performance were collected at individual GP level, practices asked to receive feedback at the level of the GP as well. *Relative standards* for determining the level of the incentive were preferred over absolute standards by most target users. Setting an absolute standard for performance was considered too arbitrary. They preferred performance standards for both indicators and subjects, but as this would not contribute to the clarity of the calculations it was decided

Table 1 Elements of the P4P program, design options and choices

Component	Elements	Design options	Design choices P4P program
Performance measurement	**Performance indicators**		
	Domains, subjects and indicators	Selection of: – Clinical care (diabetes, asthma, COPD, cardiovascular risk management, influenza vaccination, cervical cancer screening, prescribing acid suppressive drugs and antibiotics) – Practice management (infrastructure, team, information, quality and safety) – Patient experience (experience with general practitioner and organisation of care)	Selected indicators for: – Clinical care: diabetes (n=9), asthma (n=4), COPD (n=5), cardiovascular risk management (n=8), influenza vaccination (n=2), cervical cancer screening (n=1), prescribing antibiotics (n=2) – Practice management: infrastructure (n=7), team (n=8), information (n=3), quality and safety (n=4) – Patient experience: experience with general practitioner (n=16) and organisation of care (n=11)
	Period of data collection	Data collection for all three domains each year vs. a trimmed-down version of the program	At baseline measurement of clinical care, practice management, patient experience; In following years only clinical care and patient experience
Appraisal	**Unit of assessment**	Individual GP vs. general practice vs. larger organisational unit	General practice
	Performance standards	• Absolute vs. relative standards • Same standards vs. different standards for indicators/subjects	• A relative standard set at the 25th percentile of group performance • Different standards for indicators
	Analysis and interpretation of performance data		
	Weighing the domains	Different weights vs. same weight	Clinical care : practice management : patient experience 2:1:1
	Weighing the indicators	Different weights vs. same weight	Same weight for all indicators
	Calculations	• Separate scores for each domain vs. one overall domain-score • Calculations for quality level and/or improvement of performance	• Separate scores for each domain • Calculations for both quality level and improvement of performance
	Weighing the quality scores	Different weights vs. same weight for quality level and improvement of performance	Quality level : improvement of performance 3:1
	Differentation of quality scores	4 levels vs. 5-7 levels vs. 8-10 levels	7 levels

	Feedback	• Benchmark: median vs. best practice (75th or 90th percentile) vs. a combination • Risk adjustment: indirect vs. direct correction • One-step procedure vs. two-step procedure	• 25th percentile, median, 75th percentile • No risk adjustment • Two-step procedure
Reimbursement	**Financial rewards**		
	Payment	Money vs. human resources vs. sabbatical leave vs. a combination	Money
	Size of the bonus	5000 Euros to 10000 Euros (5-10% practice income) on average per practice (depending on practice size) appropriate or not?	5000 Euros to 10000 Euros on average per practice (depending on practice size) Baseline: A maximum of euro 6.89 on average per patient* Following years: A maximum of euro 2.88 on average per patient*
	Spending the bonus	No obligations vs. obligations (spending for practice with or without pre-set goal) vs. a combination	No obligations

* A patient whose health insurance company was a sponsor of the P4P program

to restrict it to the indicators only. They agreed to set the relative performance standard at the 25th percentile of each indicator. This responded to the preference of the panel members to vary the performance standards between the indicators. The target users thought that all individual indicators should receive the same *weight* because a good criterion for differentiating was lacking. However, they decided that clinical care should receive double the weight of practice management and patient experience (2:1:1) because clinical care is the major domain of quality of care. The data from the 22 practices in the field test showed that the quality scores should be calculated separately for each domain because otherwise the performance on clinical care would dominate the overall quality score. Having data available for two or more years made it possible to calculate improvement of performance as well. Panel 2 decided to reward *quality level as well as improvement of performance* in a ratio of 3:1 for the bonus payment in the next year. In that case practices with high scores would receive a bonus for delivering quality and practices with low scores would be stimulated to improve. In the following years two separate scores will be calculated for each practice; one on quality level and one on improvement of performance for the three domains clinical care, practice management and patient experience. The panel thought of the P4P program as a three years cycle in which practice management was only measured at the beginning of each cycle. The users preferred to have a detailed division in levels of quality scores to make small differences in quality visible and to make it easier to achieve next levels. The range of quality scores of all participating practices were therefore divided into *seven equal levels* (relative 'thresholds'). Practices that do not improve will be rated in level 0 (no bonus) and practices showing improvement will be rated in one of six levels with a differentiation in bonus accordingly. For feedback the target users preferred a *benchmark* with the 25th (minimum standard), 50th and the 75th percentile because that would give them a good overview and stimulate practices at the bottom as well as at the top. *Risk adjustment* for the process indicators was discussed. The target users preferred stratification, which is a comparison with a benchmark consisting of comparable practices instead of correction of their own data. However, stratification would require a large sample of practices, so we decided not to include this aspect in the experiment. Following the experiences with the field test, a *two-step procedure* was chosen by the GPs. Practices will receive feedback (indicator scores and benchmark) and the bonus after they have had the opportunity to respond to their feedback. The feedback was accompanied with clear information on the calculation procedure.

Reimbursement

The discussion on the type of financial reward resulted in the conclusion: 'Money is the best method, because money can buy you anything'. The target users

agreed on a bonus that was 5% to 10% of practice income with a minimum of 0 Euros and a maximum of 15000 Euros. None of the users indicated this amount was too high and half of them thought this was too low. However, agreement was reached with the argument that the proposed size of the bonus was in line with bonuses paid in trade and industry. A decline in the bonus for the years to follow had to do with the budget of the health insurance companies who rationalized it by arguing that the data collection was most labour intensive at the start of the project. According to the discussion on the appraisal we need two formulas to calculate the bonus, one on the quality level and one on quality improvement level. The quality improvement level can only be calculated after the first year. The formulas are:

> Bonus practice (i) on quality level (QL) = 3(2 * clinical care QL (x) + patient experience QL (y) + practice management QL (z)) * number of patients in practice insured by payers

> Bonus practice (i) on quality improvement level (QIL) = (2 * clinical care QIL (x) + patient experience QIL (y)) * number of patients in practice insured by payers

The exact bonus at baseline and in following years is presented in Table 2. The maximum bonus in year 1 is 6.890 Euros per 1000 patients (which is on average 7.500 Euros for a practice with 2350 patients) and in the following years 2.880 Euros. Target users found that formulating explicit criteria for spending the bonus was not necessary. Rewarding good quality versus penalizing poor quality was discussed in the panel as well, but proved to be not applicable at this stage of the P4P program.

Table 2 Bonus per patient for the first year and the following years for each domain and quality (improvement) level

Baseline bonus for clinical care, practice management and patient experience per patient

	Quality score	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Clinical care	Quality level	€ 0	€ 0.83	€ 1.33	€ 1.87	€ 2.37	€ 2.95	€ 3.45
Practice management	Quality level	€ 0	€ 0.41	€ 0.66	€ 0.94	€ 1.19	€ 1.47	€ 1.72
Patient experience	Quality level	€ 0	€ 0.41	€ 0.66	€ 0.94	€ 1.19	€ 1.47	€ 1.72

Bonus in following years for clinical care and patient experience per patient

	Quality score	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Clinical care	Quality level	€ 0	€ 0.25	€ 0.50	€ 0.75	€ 1.00	€ 1.25	€ 1.50
	Quality improvement	€ 0	€ 0.07	€ 0.14	€ 0.21	€ 0.28	€ 0.35	€ 0.42
Patient experience	Quality level	€ 0	€ 0.12	€ 0.25	€ 0.30	€ 0.50	€ 0.62	€ 0.75
	Quality improvement	€ 0	€ 0.03	€ 0.07	€ 0.10	€ 0.14	€ 0.17	€ 0.21

Discussion

P4P proves to be a complex innovation and knowledge needs to be acquired over time.[10] Assuming a greater probability of acceptance of P4P programs and subsequent quality improvement, our study contributes to this field by describing the design choices of target users when they themselves are involved in developing a P4P program. We succeeded in involving the target users in the lively discussions about design options. They were very much involved in the discussions and in the field test; the response rate in the panels was high. We managed to reach consensus and to define a P4P program for primary care in the Netherlands.

In line with other P4P programs our target users selected performance indicators for clinical care, practice management and patient experience. It was not surprising that the chronic diseases were chosen for the program concerning the attention for these diseases and concerning the health care costs due to these diseases. However, our program seems to be more balanced compared to other programs with regard to the position of patient experiences in the program.[3,5-7] GPs indicated that they wanted the patient to be more central in the program because patient communication is a core task. Nevertheless, they wanted clinical care to be weighted more heavily than patient experience to reduce the chance of being solely judged on patient experiences. Here the consequences of the choices seem to overrule the principles.

Mostly, P4P programs are designed and implemented top-down by policy makers and service managers.[13] In our study both GPs and health insurance companies were involved in the development of the program. Interestingly, the health insurance representatives did not want to discuss the content of clinical care and allowed the GPs to decide on this domain. In other programs the payers had a more decisive role in the development of a P4P program or were not involved in any way. Though the effectiveness of P4P programs is still inconclusive, we assume that our approach enhances the commitment and motivation of general practices and therefore the impact of our program.

The target users had a realistic estimate of the required size of the bonus in order to achieve a quality stimulus. According to the target users a bonus of on average 5% to 10% of practice income was considered to be appropriate. The target users were aware of the risk of gaming when the incentive is too high.[21,24,25] Our bonus is much smaller than the incentive in the UK which makes up approximately 25% of GPs' income.[4]

The target users opted for relative P4P standards. Until now, programs mainly base their incentives on absolute performance standards.[1,3,5,6,8,28] An advantage of relative standards is that health insurers can remain within their budget. This is in contrast to the UK P4P program, for example.[29] Furthermore quality scores of all participating practices were divided into seven levels. These series of tiered thresholds have attainable goals for each practice; a known effective stimulus

for changing behaviour.[30] According to the target users both quality level and improvement of performance need to be incentivised. This will stimulate practices with a high performance as well as practices with a low performance.[1] This is in contrast with other P4P programs in which nearly always good performance instead of improvement is rewarded.[11]

Strengths and limitations of the study

The strength of our study lies in its developmental process, assuming a greater probability of acceptance of the program and subsequent quality improvement. Involving the target users resulted in good discussions and consensus about the design options. The field test was performed successfully as part of the procedure. Many practices participated in the field test as well as in the panels, which resulted in a reasonably balanced P4P program.

This study has some limitations. First, due to time constraints patients were not included in the design of the P4P program. However, they had been previously involved in discussions about the objectives of the Dutch accreditation program, which was part of the initial framework of our program. Second, there is a drop in the number of participants in panel 2. A possible explanation is that the subjects we discussed in panel 2 were more restricted and detailed and therefore less attractive than those in panel 1. Third, practices could voluntarily register for this experiment, which may have resulted in overrepresentation of early adopters of a P4P program. It is important that the early majority as well as later on the late majority support the P4P program. To involve them in design choices that are acceptable and applicable is still a challenge.

Strengths and limitations of the design choices

A strength of the design choices is the involvement of the target users. The behavioural change of the P4P program is therefore grounded in extrinsic (reimbursement) as well as intrinsic motivation. To stimulate the motivation further the feedback will be discussed within the practice supported by a facilitator.

The performance measures do not cover all aspects of general practice. Just stimulating the incentivised parts of the performance can result in a possible decline in quality of care of the non-incentivised aspect.[31] By discussing which aspects will be stimulated in the forthcoming period, we assume that this effect is somewhat lower in our P4P program.

GPs have decided that the outcome indicators on clinical care will not be incentivised, and the health insurance companies agreed. We have to study the effect of this decision on the outcome indicators. By incentivising the process indicators an indirect effect is expected on the reported outcome measures, but that still has to be proven.

As in other P4P programs the focus of the clinical performance measures is on

the chronic conditions. Policy makers show a lot of interest in the performance on these conditions, resulting in several improvement projects. This might have an effect on our baselines measures in which case the room for improvement decreases. In our effect study we will take into account the baseline measures to get more insight into this problem.

The relative thresholds might evoke calculating behaviour, that is if no one improves the bonus will still be dived. The question is which practice will take this risk. This design choice introduces a prisoner's dilemma with unclear results. Although, based on the involvement of the participating GPs in quality of care we would assume that in this group the urge of improvement is larger. The sustainability of the relative thresholds can become tensed in a broader probably less involved group of GPs.

The health insurance companies decided together with the GPs on the available budget for the bonus. However, after the first bonus was allocated, the health insurance companies started the discussion on the bonus again. They suggested that the data collection was much easier in the following year, and therefore the bonus could be reduced. This led to a lot of turbulence among the participating GPs. No consensus was reached, but the practices were still willing to participate. This is a demonstration of the inequity in the relation between payers and GPs that is hard to cover with a consensus procedure.

The sustainability of the P4P program is also stressed if the performance measures stay the same each cycle, because the indicator scores increase due to the P4P program to a certain optimum. This phenomenon has been described with the UK-QOF data.[4] To prevent this effect it was discussed that the P4P program will need constant trimming, recalibrating and balancing to ensure that the objectives are being met at the right costs and without too many unwanted effects. This means that subjects and/or indicators will be replaced with others when performance on the subjects reaches a certain level. This will also prevent a narrow focus on quality of care in general practice. The adjustments of the P4P program should again be based on discussions with the target users.

Conclusions

By performing a procedure to involve target users in designing a P4P program for general practice, a detailed framework to define design choices was established. This framework as well as the insight into motives for design choices of the target users can be helpful for others who are developing or evaluating a P4P program. The resulting design resembled the P4P programs from other countries, but ours was also in line with target users' views and assessments of relevance and applicability. As already shown by the growing number of voluntary participants during the study, this may enhance general practitioner's commitment to the program.

References

1
Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S: Does pay-for-performance improve the quality of health care? *Ann Intern Med* 2006; 145:265-272.

2
Rosenthal MB, Frank RG: What is the empirical basis for paying for quality in health care? *Med Care Res Rev* 2006; 63:135-157.

3
BMA & NHS Employers: *Quality and Outcomes Framework guidance for GMS contract 2011/12. Delivering investment in general practice* London: BMA & NHS Employers; 2011.

4
Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M: Effects of pay for performance on the quality of primary care in England. *N Engl J Med* 2009; 361:368-378.

5
Damberg CL, Raube K, Williams T, Shortell SM: Paying for performance: implementing a statewide project in California. *Qual Manag Health Care* 2005; 14:66-79.

6
Medicare Australia: Practice Incentives Program (PIP). Medicare Australia; 2009.

7
Roland M: Linking physicians' pay to the quality of care-a major experiment in the United Kingdom. *N Engl J Med* 2004; 351:1448-54.

8
Frolich A, Talavera JA, Broadhead P, Dudley RA: A behavioral model of clinician responses to incentives to improve quality. *Health Policy* 2007; 80:179-193.

9
Maffei RM, Turner N, Dunn K: Building blocks to adopting a pay-for-performance model. *JONAS Healthc Law Ethics Regul* 2008; 10:64-69.

10
Mannion R, Davies HT: Payment for performance in health care. *BMJ* 2008; 336:306-308.

11
Rosenthal MB, Fernandopulle R, Song HR, Landon B: Paying for quality: providers' incentives for quality improvement. *Health Aff (Millwood)* 2004; 23:127-141.

12
Van Herck P, Annemans L, De SD, Remmen R, Sermeus W: Pay-for-performance step-by-step: Introduction to the MIMIQ model. *Health Policy* 2010; 102:8-17.

13
Mannion R, Davies H, Marshall M: Impact of star performance ratings in English acute hospital trusts. *J Health Serv Res Policy* 2005; 10:18-24.

14
Gruen RL, Elliott JH, Nolan ML, Lawton PD, Parkhill A, McLaren CJ, et al: Sustainability science: an integrated approach for health-programme planning. *Lancet* 2008; 372:1579-1589.

15
Scott IA: Pay for performance in health care: strategic issues for Australian experiments. *Med J Aust* 2007; 187:31-35.

16
Van Herck P, De SD, Annemans L, Remmen R, Rosenthal MB, Sermeus W: Systematic review: effects, design choices, and context of pay-for-performance in health care. *BMC Health Serv Res* 2010; 10:247.

17
Campbell SM, Braspenning J, Hutchinson A, Marshall MN: Research methods used in developing and applying quality indicators in primary care. *BMJ* 2003; 326:816-819.

18
Van Doorn AL, Kirschner K, Bouma M, Burgers JS, Braspenning JCC, Grol RPTM: Evaluatie van het onderdeel medisch handelen van de accreditering. Vier klinimetrische citeria. *Huisarts Wet* 2010; 53:141-146.

19
Grol R, Dautzenberg M, Brinkmann H, eds: *Quality Management in Primary Care. European Practice Assessment.* Gutersloh, Verlag Bertelsmann Stiftung; 2004.

20
Grol R, Wensing M, Mainz J, Ferreira P, Hearnshaw H, Hjortdahl P, et al: Patients' priorities with respect to general practice care: an international comparison. European Task Force on Patient Evaluations of General Practice (EUROPEP). *Fam Pract* 1999; 16:4-11.

21
Conrad DA, Perry L: Quality-based financial incentives in health care: can we improve quality by paying for it? *Annu Rev Public Health* 2009; 30:357-371.

22
Kiefe CI, Allison JJ, Williams OD, Person SD, Weaver MT, Weissman NW: Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. *JAMA* 2001; 285:2871-2879.

23
Rosenthal MB, Dudley RA: Pay-for-performance: will the

latest payment trend improve care? *JAMA* 2007; 297:740-744.

24
Campbell S, Reeves D, Kontopantelis E, Middleton E, Sibbald B, Roland M: Quality of primary care in England with the introduction of pay for performance. *N Engl J Med* 2007; 357:181-190.

25
Doran T, Fullwood C, Reeves D, Gravelle H, Roland M: Exclusion of patients from pay-for-performance targets by English physicians. *N Engl J Med* 2008; 359:274-284.

26
Burns D: *Systemic Action Research: A strategy for whole system change* Bristol: Policy Press; 2007.

27
Normand SL, McNeil BJ, Peterson LE, Palmer RH: Eliciting expert opinion using the Delphi technique: identifying performance indicators for cardiovascular disease. *Int J Qual Health Care* 1998; 10:247-260.

28
Conrad DA, Christianson JB: Penetrating the 'black box': Financial incentives for enhancing the quality of physician services. *Med Care Res Rev* 2004; 61:37S-68S.

29
Galvin R: Pay-for-Performance: Too Much of a Good Thing? A Conversation with Martin Roland. *Health Affairs Web Exclusive* 2006; 25: w412-w419.

30
Loewenstein G, Prelec D: Anomalies in intertemporal choice: evidence and an interpretation. *Q J Econ* 1992; 107:573-597.

31
Doran T, Kontopantelis E, Valderas JM, Campbell S, Roland M, Salisbury C, et al: Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework. *BMJ* 2011; 342:d3590.

Additional file The indicator set of P4P program

Clinical care

Diabetes

Information from the previous 12 months

1. The percentage of patients with diabetes who have had three times a glucose measurement
2. The percentage of patients with diabetes who have a record of HbA1c
3. The percentage of patients with diabetes who have a record of the blood pressure
4. The percentage of patients with diabetes who have a record of total cholesterol
5. The percentage of patients with diabetes who use cholesterol medication
6. The percentage of patients with diabetes who have a record of serum creatinine testing
7. The percentage of patients with diabetes with a record of neuropathy testing
8. The percentage of patients with diabetes who have a record of retinal screening in the previous 24 months
9. The percentage of patients with a fully completed risk profile

COPD

Information from the previous 12 months

1. The percentage of patients with COPD in whom a spirometry has been done ever
2. The percentage of patients with COPD in whom a spirometry has been done in the previous 12 months
3. The percentage of patients with COPD with which there has been contact
4. The percentage of patients with COPD in whom there is a record of smoking status
5. The percentage of patients with COPD who smoke, whose notes contain a record that smoking cessation advice has been offered

Asthma

Information from the previous 12 months

1. The percentage of patients with asthma in whom a spirometry or a peak flow measurement has been done ever
2. The percentage of patients with asthma with which there has been contact
3. The percentage of patients with asthma in whom there is a record of smoking status
4. The percentage of patients with asthma who smoke, and whose notes contain a record that smoking cessation advice has been offered

Cardiovascular risk management

Information from the previous 12 months

1. The percentage of high risk patients whose notes have a record of blood pressure
2. The percentage of high risk patients whose notes have a record of total cholesterol or cholesterol ratio
3. The percentage of high risk patients with statins
4. The percentage of high risk patients whose notes record smoking status
5. The percentage of high risk patients who smoke, and whose notes contain a record that smoking cessation advice has been offered
6. The percentage of high risk patients with a fully completed risk profile
7. The percentage of patients with heart disease in anamnesis who are using anticoagulant drugs
8. The percentage of high risk patients whose notes record a glucose measurement

Influenza vaccination

1. The percentage of vaccinated high risk patients in practice
2. The percentage of vaccinated patients of 65 years and older

Cervical cancer screening

1. Women from target cohort screening whose notes record a cervical smear

Antibiotics

Data from the previous 12 months

1 Narrow-spectrum antibiotic cures in relation to all antibiotic cure prescriptions
2 Number of antibiotic prescriptions per 1000 patients

Practice management

(answering scale yes/no)

Infrastructure

	Items
Presence of adequate space	2
Accessibility and availability	3
Presence of instruments for diagnosis and treatment	11
Presence of instruments for laboratory supplies	8
Presence of first aid facilities	11
Presence of emergency suitcase with required content	17
Presence of material for working hygienic	9

Team

	Items
Executing technical and diagnostic tasks	13
Executing organisational en administrative tasks	9
Executing tasks with regard to chronic diseases and prevention	11
Having structured internal arrangements	3
Meetings within HAGRO (general practitioner's group)	2
Having structured meetings with primary caregivers	5
Having structured meetings with health care organisations	11
Personnel policy	1

Information

	Items
Medical reporting	6
Medical and non-medical information for patients	8
Presence of social map	1

Quality and safety

	Items
Use of quality system with pharmacist	1
Working against standards and protocols	2
Quality policy	3
Education of practice employees	5

Patient experience

Patient experience with general practitioner's functioning from the previous 12 months*

Items

1 Making you feel s/he had time during consultations
2 Interest in your personal situation
3 Making it easy for you to tell him/her about your problems
4 Involving you in decisions about medical care
5 Listening to you
6 Keeping your records and data confidential

7 Quick relief of your symptoms
8 Helping you to feel well so that you can perform your normal daily activities
9 Thoroughness
10 Physical examination
11 Offering you services for preventing diseases (e.g. screening, health checks, immunizations)
12 Explaining the purpose of tests and treatments
13 Telling you what you wanted to know about your symptoms and/or illness
14 Help in dealing with emotional problems related to your health status
15 Helping you understand the importance of following his or her advice
16 Knowing what s/he had done or told you during previous contacts

* 5-point Likert scale questions and an option 'not applicable'

Patient experience with organisation of care from the previous 12 months*

Items

1 Preparing you for what to expect from specialists or hospital care
2 The helpfulness of the staff (other than the doctor)
3 Getting an appointment to suit you
4 Getting through to the practice by phone
5 Being able to speak to the GP by phone
6 Waiting time in the waiting room
7 Providing quick services for urgent health problems

Accessibility in general practice and general practitioner#

8 It is possible to ask for a longer consultation
9 General practitioner is good accessible by phone
10 Patient gets another general practitioner regularly

Procedure for complaints#

11 The practice has an accessible procedure for complaints

* 5-point Likert scale questions and an option 'not applicable'
\# Answering scale yes/no

Un-incentivised clinical outcome indicators

Diabetes

Information from the previous 12 months

1 The percentage of patients in whom the Hba1c is 7 or less
2 The percentage of patients in whom the blood pressure is 150/85 or less
3 The percentage of patients whose measured total cholesterol is 5.0 mmol/l or less

COPD

Information from the previous 12 months

1 The percentage of patients with no exacerbation

Asthma

Information from the previous 12 months

1 The percentage of patients with no exacerbation

Cardiovascular risk management

Information from the previous 12 months

1 The percentage of patients in whom the blood pressure is 160/90 or less
2 The percentage of high risk patients with statins whose measured cholesterol is 5.0 mmol/l or less

Chapter 4

Evaluation of clinical indicators. Four reliability and validity issues

Arna van Doorn
Kirsten Kirschner
Margriet Bouma
Jako Burgers
Jozé Braspenning
Richard Grol

Huisarts en Wetenschap 2010; 53(3): 141-6

Abstract

Aim To describe four reliability and validity issues regarding the clinical indicators from the Visitation Instrument Accreditation (VIA). Based on this information, practices needed to start improvement plans in order to get accreditation.

Method An observational study based on the medical records of 82 practices.

Results The indicators that covered chronic disease management (diabetes, COPD, asthma and cardiovascular risk management), prevention activities (influenza vaccination, cervical cancer screening) and antibiotics policy were correlated weakly, suggesting that the instrument provided a rather broad scope of the practice when it comes to chronic disease management and prevention. Furthermore, the different subjects were each measured by indicators that had sufficient coherence, which suggested that they measured a clear underlying concept. To achieve a reliable indicator score, data from at least 96 patients were necessary when 10% error is allowed. VIA allows to take a sample of 40 patients, but in that case the error margin increases to 15%. To establish a reliable benchmark we needed 233 practices when 5% error is allowed.

Conclusion The clinical indicators of the VIA are reliable and valid and can be used by a general practice to gain insight into their own performance compared to others. For practice policy on quality improvement an error margin of 10-15% around the indicator score on practice level seems to be acceptable. We would be more comfortable with a smaller error margin in case of accountability or a 'pay-for-performance' program. A sample of 40 patients would not do in such a case. Finding a balance between feasibility and justice therefore remains a continuous search.

Introduction

Clinical indicators are needed to measure the quality of care in order to create more transparency in healthcare.[1] To achieve this, general practices use the Visitation Instrument Accreditation (VIA), whereby several sets of indicators map the various aspects of general practice care.[2] Since 2005, the VIA has been used for the Dutch National Accreditation program of the Dutch College of General Practitioners, which consists of the following domains: clinical care, practice management and patient experience. These indicators are used to get more insight into the quality of care, and enable each general practitioner to compare the performance within their practice to other practices (benchmark). This way of giving feedback offers general practitioners clues for improving the quality of their care. This type of information is increasingly used, for instance within the framework of internal quality policies, multidisciplinary care groups and public information. Because of this, questions regarding the validity and reliability of this information are of notable importance.

The validity and reliability of the domains practice management and patient experience have been described before.[3,4] However, little is yet known about clinical care. Even so, when this set of indicators was developed, the *content validity* has been determined for the individual indicators, meaning that a panel of experts stated that the indicators are a good reflection of clinical care and can be used to evaluate this domain.[2,5] Now that data have been collected with the instrument, it is also possible to examine other aspects of reliability and validity.[6,7] This research examines clinical indicators on the basis of four reliability and validity issues.

Firstly, the various sets of indicators of the instrument should describe a broad and diverse range of clinical activities, which take place in a general practice. This means that the sets of indicators should not be too closely connected to each other (low correlation between the sets of indicators). Secondly, it is important that there is a certain amount of coherence between the indicators within one set, but the information should not overlap too much (internal consistency within a set). Thirdly, the indicator score should give an accurate picture of the practice. This means that data should be collected from a sufficient number of patients (reliable indicator score). Fourthly, in order to give a stable and correct picture, a benchmark should be based on a sufficient number of practices. Moreover, the practices on the basis of which the benchmark has been determined, should be representative of the practice for which the comparison is needed (adequate and reliable benchmark). We studied the extent to which the domain clinical care of the VIA meets each of these criteria.[2]

Method

Study population and instrument

We carried out an observational study based on the medical records of 82 practices that took part in the Dutch National Accreditation Program during 2005-2006 on a voluntary basis. The practices collected data on three widespread chronic diseases (diabetes, asthma and COPD), cardiovascular risk management, a number of specific prevention activities (influenza vaccination and cervical cancer screening) and antibiotics policy. The instrument consisted of structure indicators, process indicators and outcome indicators.[8] We mostly concentrated on the process indicators for the examination of the criteria, because these are the most reliable when it comes to giving information on the quality of practice management and care.

Box 1 Accreditation of general practices of the Dutch College of General Practitioners (NHG-Praktijkaccreditering®)

Since 2005, general practices have been using the NHG-Praktijkaccreditering® (Accreditation of general practices of the Dutch College of General Practitioners). GPs collect data for their accreditation using the Visitation Instrument Accreditation (VIA). The VIA consists of quality indicators with regard to clinical care, practice management and patient experiences, and measures the quality of (part of) the activities carried out in the general practice. Data have been extracted from medical files and questionnaires. In addition, observations have been made by a consultant of NPA Ltd, the organisation that carries out the independent review on behalf of the quality mark NHG-Praktijkaccreditering®. On the basis of the incorporated data the practice is able to compare itself to other practices, and will consequently be able to draw up plans for improvement. Accreditation of the practice is done on the basis of these plans for improvement and on their subsequent results, provided that certain conditions (minimum requirements) have been met. This will be checked by the accreditation officer at the NPA during an audit. Data will be collected using the VIA in the first year. In the following two years another audit of the practice will take place, in which the accreditation officer will check whether the practice has achieved the results aimed at, whether the plans for improvement for the upcoming year will meet the requirements and whether the minimum requirements for that particular year have been met. The cycle will repeat itself after three years.

Table 1 shows, by way of example, the nine process indicators that were used for diabetes. We used eight process indicators for cardiovascular risk management (CVRM), five for COPD, four for asthma, and one indicator each for influenza vaccination, cervical cancer screening and antibiotics policy. We also included some outcome indicators for a number of the calculations, in order to get an idea of this type of information (with regard to diabetes three and CVRM two indicators). An overview of all the indicators that were included in our study will be presented in Table 6.

Table 1 Process indicators within the subject diabetes

Indicator	Description (information from the previous 12 months)
1	The percentage of diabetes patients with a fully completed high risk profile
2	The percentage of patients with diabetes who have had three times a glucose measurement
3	The percentage of patients with diabetes who have a record of HbA1c
4	The percentage of patients with diabetes who have a record of the blood pressure
5	The percentage of patients with diabetes who have a record of total cholesterol
6	The percentage of patients with diabetes who have a record of serum creatinine testing
7	The percentage of patients with diabetes who have a record of retinal screening in the previous 24 months
8	The percentage of patients with diabetes with a record of neuropathy testing
9	The percentage of patients with diabetes who use cholesterol medication

Data collection

The data required were collected by a staff member of the general practice. The medical file constituted the source of the data. However, it was often not easy to extract data from the electronic medical records (EMR); for one thing because the required software was not available, but also because data had not been registered in a uniform way. This led to the practices being offered the choice to collect data from a sample survey of 40 patients. If the total number of patients with the disorder concerned did not exceed 40, the practices were then requested to collect details of all the patients. The VIA requested details to be collected only for that part of the patient group that is treated by the GP. However, reports were often made on all patients with the disorder concerned. With regards to the antibiotics policy data, our advice was to contact the preferential pharmacist. The data collection could refer to just one GP, or to more GPs if they shared the patient population. Of the 82 practices, we included a total of 97 patient populations; we had details on all subjects of every population.

Analysis

In order to find out whether the practices in our study were representative of all Dutch practices, we compared our study population to the Dutch general practices concerning type of practice, degree of urbanization and to what extent practices had a dispensary as well. Moreover, we looked at the number of patients per FTE GP.

In order to check whether the subjects of clinical care from the VIA differed enough from each other, we determined the level of correlation between the various subjects (criterion 1) using Pearson's correlation coefficient. To assess the internal consistency (criterion 2) we determined the correlation between the various indicators of one subject by using Cronbach's α.[9] A Cronbach's α

between 0.7 and 0.8 is advisable, although an α higher than 0.6 is also considered to be acceptable for these types of indicators.[10] Per indicator we calculated for each patient population the number of patients needed in order to achieve a reliable score (criterion 3). This calculation was based on the indicator score we found. We kept to a level of significance of 95% and tested for an error margin of both 5% and 10%. Per indicator we then calculated the maximum, the 75th percentile score, the mean and the standard deviation of the numbers found, in order to get a clear picture of the scores and the level of dispersion. We also used a power calculation to estimate the number of patient populations needed to get a reliable benchmark (criterion 4).[11] For this we kept to a level of significance of 95% and tested again for an error margin of both 5% and 10%.

Results

Study population

Most characteristics of the general practices in our study reasonably correspond to the characteristics of the Dutch general practices (Table 2). Yet there seem to be relatively fewer solo practices in our study. Also, the number of patients per FTE GP is higher in our study population than the national number of residents per fulltime equivalent (FTE) GP.

Table 2 Practice characteristics of the study population in comparison with all Dutch practices (2008)

	Study population NHG - Accreditation of GP Practices n = 82 practices*		*All Dutch general practices* n = 4,235 practices*
	n	%	%
Type of practice			
Solo practice	18	22.0	42.3
Duo practice	18	22.0	31.5
Group practice or health centre	24	29.3	26.1
HOED	15	18.3	_
Other type of practice	7	8.5	_
Degree of urbanization†			
Very strong/strong urbanization	33	40.2	46.7
Moderate/little urbanization	38	46.3	41.1
No urbanization	11	13.4	12.2
Having a dispensary			
Yes	3	3.7	7.3
No	79	96.3	92.7
Number of patients per FTE GP	2444		
Number of inhabitants per FTE GP			2322

* Data provided by NIVEL, 1-1-2008[12]; † Very strong/strong urbanization >1.500 addresses per km^2; moderate/little urbanization = 500-1.500 addresses per km^2; no urbanization <500 addresses per km^2
HOED: Construction where multiple GPs have their separate practices under one roof

Criterion 1: low correlation between the indicator sets

There were no strong correlations found between the various subjects which were dealt with within clinical care (Table 3). The subject of antibiotics did not show coherence with any of the other subjects. Although cervical cancer screening turned out to be slightly cohered to cardiovascular risk management, this correlation was rather weak. Of the other five subjects, asthma showed the strongest coherence with the other subjects (diabetes, COPD, cardiovascular risk management and influenza vaccination). Furthermore, we noticed a slight correlation between diabetes and both cardiovascular risk management and influenza vaccination.

Table 3 Correlation between the subjects within clinical care, expressed through Pearson's correlation coefficient of the means

	Diabetes	***COPD***	***Asthma***	***CVRM***	***Influenza vaccination***	***Cervical cancer screening***	***Antibiotics***
Diabetes	X	0.15	0.31*	0.43*	0,29*	0.19	-0.10
COPD		X	0.40*	0.09	0.20	-0.13	0.12
Asthma			X	0.48*	0.30*	0.16	0.06
CVRM				X	0.08	0.27*	0.19
Influenza vaccination					X	0.17	0.17
Cervical cancer screening						X	-0.07
Antibiotics							X

* Significant correlation, p <0.01; CVRM: cardiovascular risk management

Criterion 2: internal consistency of the indicator sets

Table 4 shows the average scores on the indicators for each subject, the standard deviation and the internal consistency of the indicators for each subject (Cronbach's α). The nine process indicators which relate to diabetes proved to have the highest internal consistency. These indicators had a Cronbach's α of 0.73. This score can vary between 0 and 1, whereby 0 stands for no intercorrelation at all and 1 for a perfect intercorrelation. When a score is almost 0.9 or higher, the internal consistency is so strong that there is an overlap in the information provided by the indicators. An option then would be to leave out indicators. Scores closer to 0 suggest that the indicators deal with different subjects, which means that the indicators ought not to be taken together. The eight process indicators relating to cardiovascular risk management had a reasonably strong internal consistency, just as the five indicators relating to COPD (α = 0.64 and α = 0.67 respectively). The internal consistency of the four asthma indicators proved to be slightly less strong (α = 0.56). Describing the internal consistency using Cronbach's α is only possible for subjects with more than one indicator.

Table 4 Internal consistency within each subject, expressed through Cronbach's α (n = 97)

Subject	*Number of indicators*	*Cronbach's α*	*Mean in %*	*SD in %*
Diabetes	9	0.73	72.4	11.2
COPD	5	0.67	63.1	15.4
Asthma	4	0.56	49.3	17.1
CVRM	8	0.64	49.6	11.4
Influenza vaccination	1	–	85.6	8.4
Cervical cancer screening	1	–	67.6	16.5
Antibiotics	1	–	14.3	11.5

SD: standard deviation; CVRM: cardiovascular risk management

Criterion 3: reliable indicator scores per patient population

Using the resulting indicator scores, we calculated for each patient population per indicator the number of patients needed. Table 5 lists the minimum number of patients needed per subject. When subjects consisted of more than one indicator we made our calculation for each indicator separately; Table 5 lists the lowest and the highest number of patients needed for each subject. Looking at the achieved indicator scores, the nine diabetes process indicators needed at least 363 and at most 384 patients to calculate a reliable score, with an error margin of 5%. At least 384 patients are needed for almost all subjects. An error margin of 10% means that we would need to assess 96 patients to achieve a reliable score. The average population size of the participating practices was 3,917 patients. This means that we can only calculate reliable indicator scores (with an error margin of 10%) for disorders with a minimal prevalence of 2.4%. This is the case for diabetes, asthma and cardiovascular risk management (RIVM site). The prevalence of COPD is just below 2.4%. There is a possibility not all patients are registered in general practices, which means that through an improved registration we will also be able to include enough patients for COPD, to achieve reliable indicator scores.

Table 5 Number of patients needed and number of patient populations needed for benchmark, highest and lowest number for each subject

Subject (number)	*Number of patients needed for each indicator*		*Number of populations needed for benchmark*	
	Error margin 5%	**Error margin 10%**	**Error margin 5%**	**Error margin 10%**
Diabetes (p, 9)	363-384	91-96	8-233	2-58
Diabetes (u, 3)	272-384	68-96	4-35	1-9
COPD (5)	384-384	96-96	46-146	11-37
Asthma (4)	384-384	96-96	64-175	16-44
CVRM (p, 8)	384-384	96-96	28-161	7-40
CVRM (u, 2)	384-384	96-96	39-46	10-11
Influenza vaccination (1)	372	93	11	3
Cervical cancer screening (1)	384	96	42	10
Antibiotics (1)	383	96	20	5

p = process indicators, u = result indicators; CVRM: cardiovascular risk management

Criterion 4: reliable benchmark

The number of patient populations needed to calculate a reliable benchmark depends on the dispersion of the various scores on the indicator concerned. The standard deviation per indicator varied between 5.1 for the diabetes outcome indicator 'percentage of diabetes patients with an HbA1C above 8.5' and 39.0 for the diabetes process indicator 'percentage of patients with a fully completed high risk profile.' Due to the fact that there is a strong dispersion between the various indicators, there are also large differences in the number of patient populations needed per indicator. For instance, the minimum number of patient populations needed for the diabetes indicators varies between 8 and 233 with an error margin of 5% (Table 5). To calculate a reliable benchmark for all indicators, 233 patient populations are needed when we use an error margin of 5% and 58 populations when we use an error margin of 10%. The current benchmark figures which are given to the VIA as reference material have been based on 259 populations, which is more than sufficient for this study population to maintain a maximum error margin of 5%.

Table 6 Overview of all indicators

Subject	Number	Description (information from the previous 12 months)
Diabetes	1	The percentage of diabetes patients with a fully completed high risk profile
process indicators	2	The percentage of patients with diabetes who have had three times a glucose measurement
	3	The percentage of patients with diabetes who have a record of HbA1c
	4	The percentage of patients with diabetes who have a record of the blood pressure
	5	The percentage of patients with diabetes who have a record of total cholesterol
	6	The percentage of patients with diabetes who have a record of serum creatinine testing
	7	The percentage of patients with diabetes who have a record of retinal screening in the previous 24 months
	8	The percentage of patients with diabetes with a record of neuropathy testing
	9	The percentage of patients with diabetes who use cholesterol medication
outcome indicators	1	The percentage of patients with diabetes in whom the Hba1c is 8.5 or more
	2	The percentage of patients in whom the blood pressure is 150/85 or less
	3	The percentage of patients whose measured total cholesterol is 5.0 mmol/l or less
COPD	1	The percentage of patients with COPD in whom a spirometry has been done ever
	2	The percentage of patients with COPD in whom a spirometry has been done in the previous 12 months
	3	The percentage of patients with COPD with which there has been contact
	4	The percentage of patients with COPD in whom there is a record of smoking status
	5	The percentage of patients with COPD who smoke, whose notes contain a record that smoking cessation advice has been offered

Asthma	1	The percentage of patients with asthma in whom a spirometry or a peakflow measurement has been done ever
	2	The percentage of patients with asthma with which there has been contact
	3	The percentage of patients with asthma in whom there is a record of smoking status
	4	The percentage of patients with asthma who smoke, and whose notes contain a record that smoking cessation advice has been offered
CVRM process indicators	1	The percentage of high risk patients whose notes have a record of blood pressure
	2	The percentage of high risk patients whose notes have a record of total cholesterol or cholesterol ratio
	3	The percentage of high risk patients with statins
	4	The percentage of high risk patients whose notes record smoking status
	5	The percentage of high risk patients who smoke, and whose notes contain a record that smoking cessation advice has been offered
	6	The percentage of high risk patients with a fully completed risk profile
	7	The percentage of patients with heart disease in anamnesis who are using anticoagulant drugs
	8	The percentage of high risk patients whose notes record a glucose measurement
outcome indicators	1	The percentage of patients with CVRM in whom the blood pressure is 160/90 or less
	2	The percentage of high risk patients with statins whose measured cholesterol is 5.0 mmol/l or less
Influenza vaccination	1	Percentage of vaccinated high risk patients in the practice or percentage of vaccinated patients of 65 years and older*
Cervical cancer screening	1	Percentage of women from the target cohort whose notes record a cervical smear
Antibiotics	1	Percentage of narrow-spectrum antibiotic cures in relation to all antibiotic cure prescriptions

* We decided to use the highest value of these two indicators to serve as indicator score for this item, because practices were often not able to fill in both indicators

Discussion

The results show that clinical care, being part of the VIA, paints a relatively broad and diverse picture of general practice care. The correlations between the subjects that have been included therein are relatively weak. The correlations we did find can be explained by the overlap between the various groups. For example, a patient who had been included in the indicators concerning cardiovascular risk management could also suffer from diabetes. Such a patient would therefore also qualify for an influenza vaccination. More subjects can be added to the instrument in the future, making the evaluation more complete and covering the field of general practice care in a broader way. The subjects that are discussed in the present version are assessed by means of a set of indicators, which show a reasonable internal consistency. The exception was the Cronbach's α score relating

to asthma, which showed a rather low coherence. A possible explanation for a lower coherence between asthma indicators is the fact that the checkup policy is maintained on a less stricter basis for asthma, especially when the patient is on little or no medication. A reasonable to good coherence between indicators serves as an indication that we are measuring one underlying concept. Often the same maximum number of patients needed emerged with the various indicators with respect to establishing reliable scores. The reason for this was that with almost all indicators at least one patient population had a score of 50%; with this score the largest number of patients is statistically needed for a reliable score. The number of patients needed decreases the closer a score gets to 0% or 100%. The reliability of the scores is also dependent on the reliability of the data on the basis of which these scores have been calculated. The scores will never be reliable when the data have not been collected in a reliable way, for instance because of problems with the data collection. Uniform reporting is therefore of the utmost importance. Moreover, the data that will be included or excluded should be determined in advance.

The question of how many patients are needed to arrive at a reliable score and which level of error would be acceptable, depends to a large extent on the purpose of the measurement. The score does not have to be accurate to one percent in order to give a good picture of possible points for improvement if it is used for internal quality improvements only. Drawing a line with respect to this is rather arbitrary and partly determined by feasibility factors. A general practice only has a limited number of patients with a certain condition. There is also a limit to the time investment by the general practitioner. The sample survey of 40 patients which is permitted at the moment by the VIA is reliable as long as an error margin of 15% is accepted. In the absence of large amounts of data, these scores could indeed be used as a reasonable indication for internal quality improvement. Stricter demands are put upon the instrument when it should be able to differentiate between practice scores or when it should report scores more accurately. If the instrument is used to distinguish between practices, for instance as part of a pay-for-performance-system, larger numbers of patients will have to be included. An error margin of 5% or less is advisable for this purpose, which means a distinction can only be made on the level of the care group or between larger practices.

In our study we did not calculate indicator scores on a practice level, but on the level of the patient population. In most cases a practice contained only one population. When practices supply data for several populations, these data can be taken together in order to give a better picture of the practice as a whole. When interpreting data it should be taken into consideration that the division between 'reliable' and 'not reliable' is not as clear as it might seem. Even if, on the basis of the number of patients included, the conclusion is that an indicator

score is not reliable, such a score could still give a reasonably good picture of the practice. However, we are not 95% sure that the score is indeed representative, meaning that care must be taken when conclusions are drawn on the basis of the scores found. The number of patients or patient populations needed will decrease by raising the error margin to 15%. Depending on the purpose, the error margin chosen and the number of patients included can be varied upon in order to arrive at a score which is as reliable as possible.

Our study shows that clinical care, being part of the VIA, is reliable and valid with regard to the four criteria studied. However, it is important in this respect to take an acceptable error margin into account when interpreting data. Uniform reporting is of the utmost importance to improve the reliability of the data collection. It has proved itself to be a valuable instrument for the GP to get more insight into his own way of acting, compared with national reference figures. On the basis of this instrument, GPs can formulate plans to improve the quality of clinical care. The instrument can also be useful when looking at the quality of clinical care on a national scale. However, the number of patients on whom data should be collected would then need to be larger than for use of the instrument within the practice, due to the higher demands on reliability.

References

1
Grol R, Wensing M. *Implementatie; effectieve verbetering van de patiëntenzorg*. 3e druk. Maarssen: Elsevier gezondheidszorg, 2006.

2
Braspenning J, Dijkstra R, Tacken M, Bouma M, Witmer H. *Visitatie instrument accreditering (VIA®)*. Nijmegen/Utrecht: WOK/NHG/NPA, 2007.

3
Grol R, Wensing M, Mainz J, Ferreira P, Hearnshaw H, Hjortdahl P, et al. Patients' priorities with respect to general practice care: an international comparison. European Task Force on Patient Evaluations of General Practice (EUROPEP). *Fam Pract* 1999;16:4-11.

4
Van den Hombergh P, Grol R, Van den Hoogen HJ, Van den Bosch WJ. Assessment of management in general practice: validation of a practice visit method. *Br J Gen Pract* 1998;48:1743-50.

5
Braspenning JCC, Pijnenborg L, In 't Veld CJ, Grol RPTM, redactie. *Werken aan kwaliteit in de huisartsenpraktijk. Indicatoren gebaseerd op de NHG-Standaarden*. Houten: Bohn Stafleu van Loghum, 2005.

6
Campbell S, Braspenning J, Hutchinson A, Marshall M. Research methods used in developing and applying quality indicators in primary care. *Qual Saf Health Care* 2002;11:358-64.

7
Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 3e druk. Oxford: Oxford University Press, 2003.

8
Donabedian A. *Explorations in quality assessment and monitoring. The definition of quality and approaches to its assessment.* Michigan: Health Administration Press, 1980.

9
Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297-334.

10
Bland JM, Altman DG. Statistics notes: Cronbach's alpha. *BMJ* 1997;314:572.

11
Moore DS, McCabe GP. *Statistiek in de praktijk*; Schoonhoven: Academic Service, 1999.

12
Hingstman L, Kenens RJ. *Cijfers uit de registratie van huisartsen – peiling 2008*. Utrecht: NIVEL, 2009.

Chapter 5

Assessment of a pay-for-performance program in primary care designed by target users

Kirsten Kirschner
Jozé Braspenning
Reinier P Akkermans
JE Annelies Jacobs
Richard Grol

Family Practice 2013; 30(2): 161-71

Abstract

Background Evidence for pay-for-performance (P4P) has been searched for in the last decade as financial incentives increased to influence behaviour of health care professionals to improve quality of care. The effectiveness of P4P is inconclusive, though some reviews reported significant effects.

Objective To assess changes in performance after introducing a participatory P4P program.

Design An observational study with a pre- and post-measurement.

Setting and subjects Sixty-five general practices in the south of the Netherlands.

Intervention A P4P program designed by target users containing indicators for chronic care, prevention, practice management and patient experience (general practitioner's [GP] functioning and organisation of care). Quality indicators were calculated for each practice. A bonus with a maximum of 6890 Euros per 1000 patients was determined by comparing practice performance with a benchmark.

Main outcome measures Quality indicators for clinical care (process and outcome) and patient experience.

Results We included 60 practices. After 1 year, significant improvement was shown for the process indicators for all chronic conditions ranging from +7.9% improvement for cardiovascular risk management to +11.5% for asthma. Five outcome indicators significantly improved as well as patients' experiences with GP's functioning and organisation of care. No significant improvements were seen for influenza vaccination rate and the cervical cancer screening uptake. The clinical process and outcome indicators, as well as patient experience indicators were affected by baseline measures. Smaller practices showed more improvement.

Conclusions A participatory P4P program might stimulate quality improvement in clinical care and improve patient experiences with GP's functioning and the organisation of care.

Introduction

The effectiveness of pay-for-performance (P4P) programs is still inconclusive,[1,2] despite the proliferation of these programs. Reviews give some suggestions for possible successful elements.[1,3] One of these might be the thorough and direct involvement of target users throughout development, implementation and evaluation phases.[4] In four studies that involved target users, three different P4P programs were evaluated by using 3-11 clinical indicators, reporting an average improvement of 20% over 3-5 years.[5-8] The studies mentioned no details of the decision process concerning the design choices of the P4P program in terms of performance measures, appraisal and reimbursement. To contribute to the knowledge of the effectiveness of P4P programs involving target users, we initiated an experiment. General practitioners (GPs) and deputies from two financing health insurance companies were invited to reach consensus on the design choices of a P4P model based on a summary of available literature on the effectiveness of P4P programs.[9] The target users took into account five out of the seven lessons from behavioural economics that might enhance the effectiveness of P4P programs.[10] First, to enhance the psychological motivation by introducing smaller and more frequent incentives instead of a large single lump-sum incentive,[11] performance on clinical indicators and on practice management and patient experience were rewarded separately. Second, a series of tiered thresholds was introduced to have attainable goals for each practice.[12] Third, the payment was realized in relatively short time, 4 months after the data collection, to highlight the value of the money (hyperbolic discounting).[13] Fourth, the payment was disconnected from usual reimbursement to get a more effective mental accounting of the incentive.[14] Fifth, as objectives and services may be a stronger driver for behavioural change than money, these types of incentives were considered.[14] The two other lessons from behavioural economics were discussed, but target users chose different options. A bonus was chosen instead of a possible more effective withhold, but in behavioural economics itself, it is unclear if the stimulus of the withhold outweighs the negative psychological reaction to unfairness.[15] Furthermore, relative instead of absolute thresholds were chosen although these might provoke uncertainty and complexity that can negatively influence the effectiveness of a P4P program.[10,16] The design choices are described in Box 1.

After the design was completed, the P4P program was implemented in the general practices. During the implementation phase, intensive contact with the target users was continued to evaluate facilitators and barriers of the program. In this paper, we describe the influence of the participatory P4P program on clinical indicators as well as on patient experience. In an observational study, we compared baseline measurement with performance after 1 year. In addition, we examined whether the change in performance differed between different types

of practices.[17-19] In the UK, group practices delivered higher Quality and Outcome Framework (QOF) scores.[17] Furthermore, Ashworth et al. found smaller practices more likely to be 'poor performers',[18] but better in access to care.[19] Preventive care proved to be worse in practices located in areas with lower socioeconomic status.[18] We expect these factors to be of influence on quality improvement and will explore their impact.

Box 1 Design choices of the participatory P4P program

Performance measurement

- ***Data collection for clinical care, practice management and patient experience***
  - Clinical care: diabetes (n = 9 indicators), COPD (n = 5 indicators), asthma (n = 4 indicators), cardiovascular risk management (n = 8 indicators), influenza vaccination (n = 2 indicators), cervical cancer screening (n = 1 indicator)
  - Practice management (n = 4 indicators): infrastructure (n = 7 items), team (n = 8 items), information (n = 3 items), quality and safety (n = 4 items)
  - Patient experience (n = 2 indicators): experience with general practitioner (n = 16 items) and organisation of care (n = 11 items)

Appraisal

- Separate appraisal of clinical care, practice management and patient experience, weighting the domains as 2:1:1
- A benchmark with relative standards was set at the 25th percentile of group performance
- For the appraisal a series of tiered thresholds was used (7 levels)
- Practices received short-term feedback (4 months after data collection)
- Both quality level and the improvement of performance were valued, with these levels weighted as 3:1

Reimbursement

- A bonus of 5% to 10% of the practice income, not linked to the usual reimbursement
- Bonus was paid in money, and not in goods or services
- Bonus to spend freely

Methods

Study design and population

An observational study with a pre- and post-measurement was conducted in a group of general practices in the south of the Netherlands, which also participated in deciding on the design of the P4P program.

Intervention

The general practices participated in designing the P4P program, see Box 1,[19] and its implementation. In the program, the participating general practices had to collect data for a set of quality indicators that described their performance in the past year in the areas of clinical care, practice management and patient experience. The indicators were systematically developed based on evidence-based guidelines and international literature. Clinical data were collected for diabetes, chronic obstructive pulmonary disease (COPD), asthma, cardiovascular risk management, influenza vaccinations and cervical cancer screening by

extracting data from electronic medical records. If routine extraction was not possible, general practices could take a random sample of 40 patients for each chronic condition. Reports from secondary databases could be used for information on influenza vaccination and cervical cancer screening. Practice management data were collected through questionnaires filled in by the practice manager. These data were checked by an independent consultant as part of the Visitation Instrument to assess practice management.[20]

To assess patient experiences of organisation of care, each practice had to distribute 40 questionnaires randomly. To assess patient experiences of the GP's functioning, each individual GP had to distribute 40 questionnaires as well. The participants were given 3 months to collect their data. Based on these data, the research team calculated quality indicators for each practice. The general practices received feedback and were given the opportunity to respond to the feedback. Four months after submitting the data, the practices received a bonus according to their performance, based on comparing their quality indicator scores of clinical care, practice management and patient experience separately with the relative thresholds, that is the 25th percentile of the average indicator scores of all the participants. For clinical care only the process indicators were incentivised.

The quality scores for clinical care, practice management and patient experience were divided into seven levels (tiered thresholds). A practice with a quality score in the lowest group did not receive a bonus. The payment was paid per patient and grew proportionally for each extra level, see Table 1. The maximum bonus was about 5% to 10% of the practice income, that is, about 6890 Euros per 1000 patients.

Table 1 Bonus for clinical care, practice management and patient experience per 1000 patients

	Quality score	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Clinical care	Quality level	€ 0	€ 830	€ 1330	€ 1870	€ 2370	€ 2950	€ 3450
Practice organisation	Quality level	€ 0	€ 410	€ 660	€ 940	€ 1190	€ 1470	€ 1720
Patient experience	Quality level	€ 0	€ 410	€ 660	€ 940	€ 1190	€ 1470	€ 1720

Variables and measurements

Clinical process indicators were calculated for diabetes (n = 9), COPD (n = 5), asthma (n = 4), cardiovascular risk management (n = 8), influenza vaccination (n = 2) and cervical cancer screening (n = 1). Clinical outcome indicators were collected for diabetes (n = 3), COPD (n = 1), asthma (n = 1) and cardiovascular risk management (n = 2). The indicators were calculated by dividing the nominator (frequency of recommended process per patient) by the denominator (number of target patients) and were expressed in a percentage between 0 and

100. The dichotomous practice management scores were summarized at practice level and presented as percentages of the total maximum possible score.[21] Patient experience was measured with the internationally validated EUROPEP instrument[22] and evaluated the GP (16 items) as well as the organisation of care (11 items). Patients scored the responses on a five-point Likert scale ranging from 'very poor' to 'excellent' and a category 'not applicable'. To calculate the item scores for patient experience, we used the percentage of patients who used the two most positive answering categories (four and five) of all patients who answered the question (other than 'not applicable').[23] The indicator scores were calculated by averaging the item scores for the evaluation of the GP and the organisation of care separately. For each practice, data on practice type (solo, duo and group) and urbanization level (large city, small city and rural area) were collected. A year after the baseline measurements, the general practices were again asked to collect data pertaining to clinical care and patient experiences. Data for practice management were not collected again due to the workload associated with this data collection.

Analysis

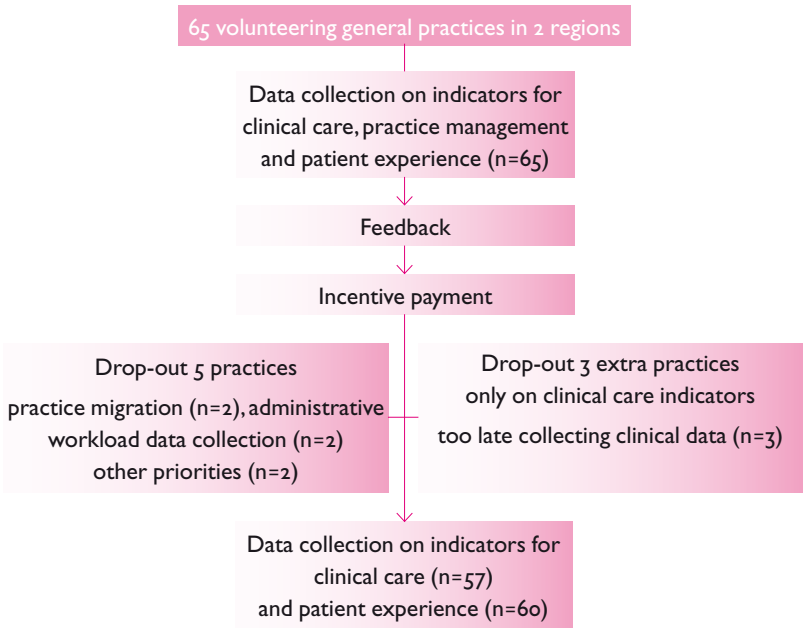
Descriptive analysis (mean, range and standard deviations) were used to summarize the indicator scores. For the chronic conditions such as diabetes, COPD, asthma and cardiovascular risk management, mean sum scores were calculated by adding the process indicators for each condition and dividing them by the number of indicators. Mean sum scores were only calculated for the practices with no missing values on all indicators for a specific chronic condition. In the analyses, the dependent variables were the clinical process and outcome indicators for diabetes, COPD, asthma, cardiovascular risk management, influenza vaccination and cervical cancer screening, the mean sum scores of the four chronic conditions, the items of patient experience individually and the two indicators of patient experience. The covariates were the baseline measurement (centralized for mean performance in year 1 of all practices), practice type (solo, duo and group) and urbanization level (rural area, small city and large city). Because of the hierarchical structure of our study (repeated measurements nested within practice), the analyses were based on a linear mixed effect model, with a random intercept and all other variables fixed. First, mixed effect models were conducted to assess the differences in change between the performance in year 1 and 2 for all clinical process and outcome indicators and for the items of patient experience with GP's functioning and organisation of care. Second, we analysed whether the differences in change were caused by the above mentioned covariates by including a covariate by effect interaction term in the model. We will only present the significant interactions and their main effect terms. All analyses were conducted using SPSS 18.0.

Results

Study population

Sixty general practices in the south of the Netherlands participated in the evaluation of the P4P program. Solo practices were underrepresented as were practices in large cities, see Table 2. Figure 1 shows the number of practices that collected data in year 1 and 2.

Figure 1 Flow diagram of participating general practices and data collection in the P4P program



Improvement on incentivised indicators

Significant improvements were shown on all clinical process indicators for diabetes care, four out of five indicators for COPD care, three out of four indicators for asthma care and three out of eight indicators for cardiovascular risk management, see Table 3. The improvements ranged from +4.2% to +26.3%. No improvements of influenza vaccination rate or the uptake rate in cervical cancer screening were observed.

The differences in change were affected by baseline measures, that is, higher baseline scores were associated with lower improvement scores (Table 4). The effects of practice type and urbanization level on the improvement are shown in Table 4 and will be explained for the COPD process indicators in detail. Group practices with an average baseline score showed a mean decline of 0.9%. The zero coefficients for solo and duo practice showed that the baseline values were equal for all types of practices. Furthermore, for every percentage point that

Table 2 Practice characteristics of the study population in comparison with all Dutch general practices

	Study population 60 practices		*All Dutch general practices*[*] *4235 practices*	
	Number	**%**	**Number**	**%**
Practice type				
Solo	16	26.7		42.3
Duo	20	33.3		31.5
Group / health centre	24	40.0		26.1
Urban area				
Large city, >1500 addresses per km²	17	28.3		46.7
Small city, 500-1500 addresses per km²	36	60.0		41.1
Rural area, <500 addresses per km²	7	11.7		12.2
Practice size				
Mean patient population per practice	4685			
Mean patient population per full-time equivalent general practitioner	2470			
Mean population per full-time equivalent general practitioner			2322	

* Since 1/1/2008. Source: NIVEL[38]

a practice scored higher than the average baseline score, the practice showed an extra decline of 0.4%. Solo practices improved 15.4% more and duo practices 14.0% more than group practices. On the asthma indicators, solo practices improved 18.3% more and duo practices improved 15.8% more than group practices. Urbanization level had no influence on the differences in change between year 1 and 2.

Patients were very positive about GP's functioning and organisation of care in year 1, nevertheless they were significantly more positive in year 2 (Table 5). Looking in more detail, improvements occurred on 26 of 27 items. The improvements in patient experience were affected by baseline measures, that is, higher baseline scores were associated with lower improvement scores (Table 4). For organisation of care, group practices with an average baseline score showed a mean improvement of 3.1%. Solo practices improved 5.2% more than group practices.

Improvement on non-incentivised clinical outcome indicators

Five out of seven outcome indicators showed significant improvements, ranging from +5.9% to +14.7% (Table 3). Baseline measures had a significant effect on these improvements, that is, higher baseline scores were related to lower improvement scores. Practices in large cities improved 14.4% less than practices in rural areas on the diabetes outcome indicator on HbA1c (Table 4). Solo practices improved 15.5% and 14.4% more on the cardiovascular risk management outcome indicators of blood pressure and cholesterol than group practices.

Table 3 Performance in year 1 and 2 and the improvement on clinical process and outcome indicators

Condition	Indicator	Indicator type	Year 1		Year 2		Improvement
			Mean score (range, SD) %	*No. of practices*	*Mean score (range, SD) %*	*No. of practices*	%
Diabetes	Three times glucose measured	Process	84.7 (27.0-100) (SD 15.0)	57	89.6 (63.5-100) (SD 10.5)	56	4.7*
	HbA1c measured	Process	91.0 (47.5-100) (SD 10.9)	57	96.2 (64.0-100) (SD 5.8)	57	5.2*
	Blood pressure measured	Process	92.8 (24.0-100) (SD 11.7)	57	97.0 (84.9-100) (SD 3.6)	57	4.2*
	Total cholesterol measured	Process	84.9 (40.0-100) (SD 12.3)	57	92.5 (38.6-100) (SD 10.0)	57	7.6*
	Use of cholesterol medication	Process	62.2 (29.9-93.3) (SD 15.5)	43	70.3 (27.1-96.0) (SD 14.9)	55	8.1*
	Creatinine tested	Process	87.2 (50.0-100) (SD 10.1)	57	93.7 (63.5-100) (SD 7.5)	57	6.5*
	Neuropathy testing performed	Process	65.4 (2.9-100) (SD 15.5)	54	78.8 (1.8-100) (SD 24.4)	57	14.0*
	Retinal screening performed in the previous 24 months	Process	72.4 (34.3-100) SD (16.9)	55	81.4 (29.4-100) (SD 16.5)	57	9.2*
	Fully completed risk profile	Process	48.2 (0-100) (SD 39.7)	52	74.2 (0-100) (SD 33.9)	57	26.3*
	Mean score diabetes (9 process indicators)		75.7 (50.0- 91.2) (SD 10.0)	39	86.0 (56.0-97.9) (SD 9.5)	54	10.4*
	Hba1c controlled (<7.0%)	Outcome#	61.4 (0-100) (SD 16.6)	57	69.0 (37.7-91.2) (SD 12.1)	57	7.7*
	Blood pressure controlled (<150/85 mm Hg)	Outcome#	62.7 (23.5-87.1) (SD 14.7)	56	68.5 (7.3-98.1) (SD 18.0)	56	5.9*
	Total cholesterol controlled (<5.0 mmol/l)	Outcome#	61.3 (0-97.5) (SD 16.8)	57	70.1 (22.0-94.0) (SD 15.0)	57	8.8*
COPD	Spirometry testing performed ever	Process	74.2 (0-100) (SD 22.4)	57	84.6 (0-100) (SD 20.5)	56	10.0*
	Spirometry testing performed in the previous 12 months	Process	51.7 (0-100) (SD 27.6)	55	57.0 (6.9-100) (SD 23.9)	55	5.1
	Contact with patient	Process	70.5 (27.0-100) (SD 20.6)	50	79.5 (31.0-100) (SD 15.9)	56	8.9*
	Smoking status recorded	Process	72.0 (0-100) (SD 25.3)	57	84.9 (11.1-100) (SD 20.5)	56	12.7*
	Smoking cessation advice offered	Process	58.7 (0-100) (SD 34.5)	53	72.7 (0-100) (SD 28.4)	56	13.9*
	Mean score COPD (5 process indicators)		67.3 (23.2-98.9) (SD 17.1)	45	76.5 (17.2-100) (SD 16.3)	55	8.1*
	No exacerbation	Outcome#	72.0 (2.3-100) (SD 19.8)	57	74.4 (3.5-100) (SD 21.0)	55	2.5
Asthma	Spirometry/peakflow testing ever	Process	53.1 (0-100) (SD 25.5)	56	65.8 (0-100) (SD 26.6)	56	12.6*
	Contact with patient	Process	58.2 (0-100) (SD 23.1)	50	65.5 (22.9-100) (SD 21.1)	56	7.3
	Smoking status recorded	Process	57.5 (0-100) (SD 31.2)	55	72.6 (8.6-100) (SD 28.0)	56	15.3*
	Smoking cessation advice offered	Process	43.8 (0-100) (SD 40.8)	50	65.0 (0-100) (SD 35.0)	54	21.1*

	Mean score asthma (4 process indicators)		55.9 (0-100) (SD 21.7)	44	67.9 (24.3-100) (SD 21.2)	54	11.5*
	No exacerbation	Outcome#	85.3 (0-100) (SD 19.5)	57	89.7 (59.1-100) (SD 10.1)	56	4.4
CVRM	Blood pressure measured	Process	79.8 (24.4-100) (SD 13.3)	56	83.4 (48.6-100) (SD 13.3)	56	3.6
	Total cholesterol or cholesterol ratio measured	Process	63.6 (19.1-97.5) (SD 17.4)	56	75.1 (20.1-100) (SD 17.1)	56	11.5*
	Statins prescribed	Process	44.8 (0-72.5) (SD 15.7)	50	48.9 (13.9-85.0) (SD 16.1)	57	4.3
	Smoking status recorded	Process	63.8 (0-100) (SD 27.4)	46	69.1 (17.1-100) (SD 26.5)	56	5.4
	Smoking cessation advice offered	Process	58.9 (0-100) (SD 38.8)	42	66.3 (1.1-100) (SD 30.0)	51	8.5
	Fully completed risk profile	Process	18.7 (0-100) (SD 25.7)	54	43.8 (0-100) (SD 32.7)	55	25.0*
	Anticoagulant drugs prescribed for patients with heart disease in anamnesis	Process	46.4 (0-95.9) (SD 24.6)	47	46.9 (7.3-98.7) (SD 23.5)	57	0.7
	Glucose measured	Process	65.2 (22.8-100) (SD 15.8)	56	72.4 (20.0-100) (SD 18.5)	57	7.2*
	Mean score Cardiovascular risk management (8 process indicators)		55.2 (37.8-86.6) (SD 10.8)	38	63.3 (30.0-92.9) (SD 14.2)	49	7.9*
	Blood pressure controlled (< 160/90 mm Hg)	Outcome#	55.7 (0-98.3) (SD 17.4)	57	70.6 (10.1-98.0) (SD 16.5)	54	14.7*
	Cholesterol controlled of patients with statins (< 5.0 mmol/l)	Outcome#	43.8 (11.1-83.3) (SD 17.3)	42	52.8 (6.0-86.6) (SD 17.3)	54	8.4*
Influenza	High risk patients vaccinated	Process	78.5 (0-97.0) (SD 15.4)	51	77.0 (42.3-96.7) (SD 11.0)	57	-1.2
	Patients of >65 years vaccinated	Process	79.5 (0-100) (SD 19.1)	51	81.5 (12.8-96.9) (SD 12.8)	57	2.0
Cervix	Cervical smear record for women from target cohort screening	Process	71.9 (0-97.7) (SD 16.8)	54	72.5 (37.8-95.6) (SD 14.6)	56	0.6

* $p < 0.05$; SD: standard deviation; # The outcome indicators were not incentivised

Table 4 Mixed models analysis for clinical process, outcome and patient experience

Clinical process	**Diabetes (n = 9 indicators) %**		**COPD (n = 5 indicators) %**		**Asthma (n = 4 indicators) %**		**CVRM (n = 8 indicators) %**	
	B (SE)	**P-value**	**B (SE)**	**P-value**	**B (SE)**	**P-value**	**B(SE)**	**P-value**
Change	10.9 (1.4)	<0.01	-0.9 (2.8)	0.75	0.7 (4.5)	0.87	7.7 (2.2)	<0.01
Baseline	1.0 (0.1)	<0.01	1.0 (0.1)	<0.01	1.0 (0.1)	<0.01	1.0 (0.1)	<0.01
Practice type#: Solo			0 (3.4)	1.0	0 (5.1)	1.0		
Duo			0 (3.0)	1.0	0 (4.4)	1.0		
Change*baseline	-0.7 (0.1)	<0.01	-0.4 (0.1)	<0.01	-0.7 (0.1)	<0.01	-0.3 (0.2)	0.17
Change*solo practice#			15.4 (4.9)	<0.01	18.3 (7.4)	0.02*		
Change*duo practice#			14.0 (4.2)	<0.01	15.8 (6.2)	0.02*		

Clinical outcome	**Diabetes Hba1c <7.0%**		**Diabetes blood pressure %**		**Diabetes cholesterol <5.0 %**		**COPD no exacerbation %**	
	B (SE)	**P-value**	**B (SE)**	**P-value**	**B (SE)**	**P-value**	**B(SE)**	**P-value**
Change	16.7 (4.1)	<0.01	5.8 (2.3)	0.02*	8.8 (1.9)	<0.01	2.4 (2.8)	0.39
Baseline	1.0 (0.1)	<0.01	1.0 (0.1)	<0.01	1.0 (0.1)	<0.01	1.0 (0.1)	<0.01
Urbanization##: Large city	0 (3.5)	1.0						
Small city	0 (3.2)	1.0						
Change*baseline	-0.7 (0.1)	<0.01	-0.6 (2.3)	<0.01	-0.7 (0.1)	<0.01	-1.0 (0.1)	<0.01
Change*large city##	-14.4 (4.9)	0.01*						
Change*small city##	-8.3 (4.6)	0.08						

Clinical outcome	**Asthma no exacerbation %**		**CVRM blood pressure %**		**CVRM cholesterol <5.0 %**	
	B (SE)	**P-value**	**B(SE)**	**P-value**	**B (SE)**	**P-value**
Change	4.4 (1.3)	<0.01	7.3 (3.3	0.03*	3.0 (4.0)	0.46
Baseline	1.0 (0.0)	<0.01	1.0 (0.1)	<0.01	1.0 (0.1)	<0.01
Practice type#: Solo			0 (3.6)	1.0	0 (4.6)	1.0
Duo			0 (3.3)	1.0	0 (3.9)	1.0
Change*baseline	-1.0 (0.1)	<0.01	-0.8 (0.1)	<0.01	-0.7 (0.1)	<0.01
Change*solo practice#			15.5 (5.2)	<0.01	14.4 (6.5)	0.03*
Change*duo practice#			9.0 (4.7)	0.06	2.8 (5.6)	0.62

Patient experience	**GP's functioning, 16 items %**		**Organisation of care, 11 items %**	
Change	**B (SE)**	**P-value**	**B (SE)**	**P-value**
Baseline	5.9 (0.5)	<0.01	3.1 (1.4)	0.03*
Practice type#: Solo	1.0 (0.1)	<0.01	1.0 (0.1)	<0.01
Duo			0 (1.6)	1.0
Change*baseline			0 (1.5)	1.0
Change*solo practice#	-0.6 (0.1)	<0.01	-0.6 (0.1)	<0.01
Change*duo practice#			5.2 (2.3)	0.03*
			3.5 (2.1)	0.09

* p <0.05; # reference group is group practice; ## reference is rural area

Table 5 Performance in year 1 and 2 and the improvement on patient experience with GP's functioning and organisation of care

	Items	Year 1	Year 2	Improvement
		Mean score	*Mean score*	
GP's functioning	Making you feel s/he had time during consultations	85.6 (SD 7.8)	92.9 (SD 5.5)	7.3*
	Interest in your personal situation	83.0 (SD 7.4)	89.0 (SD 7.7)	6.0*
	Making it easy for you to tell him/her about your problems	85.6 (SD 7.1)	92.0 (SD 5.1)	6.4*
	Involving you in decisions about medical care	82.5 (SD 7.7)	91.0 (SD 5.9)	8.5*
	Listening to you	89.4 (SD 6.9)	92.9 (SD 5.5)	3.4*
	Keeping your records and data confidential	92.9 (SD 4.3)	95.5 (SD 3.2)	2.6*
	Quick relief of your symptoms	75.9 (SD 9.2)	82.1 (SD 6.8)	6.3*
	Helping you to feel well so that you can perform your normal daily activities	79.9 (SD 8.5)	87.7 (SD 5.9)	7.8*
	Thoroughness	83.9 (SD 7.7)	90.1 (SD 5.5)	6.2*
	Physical examination	86.6 (SD 6.7)	91.4 (SD 5.5)	4.8*
	Offering you services for preventing diseases (e.g. screening, health checks, immunizations)	78.9 (SD 10.2)	84.1 (SD 7.9)	5.1*
	Explaining the purpose of tests and treatments	85.6 (SD 7.6)	91.7 (SD 5.4)	6.0*
	Telling you what you wanted to know about your symptoms and/or illness	86.2 (SD 7.6)	91.1 (SD 5.1)	4.9*
	Help in dealing with emotional problems related to your health status	78.3 (SD 10.9)	84.8 (SD 9.2)	6.5*
	Helping you understand the importance of following his or her advice	81.3 (SD 8.1)	88.4 (SD 6.8)	7.1*
	Knowing what s/he had done or told you during previous contacts	79.8 (SD 9.2)	85.5 (SD 7.3)	5.7*
	Mean score General practitioner's functioning (16 items)	83.5 (SD 6.7)	89.4 (SD 4.8)	5.9*
Organisation of care	Preparing you for what to expect from specialists or hospital care	76.6 (SD 11.5)	85.7 (SD 7.5)	9.2*
	The helpfulness of the staff (other than the doctor)	84.5 (SD 8.6)	89.3 (SD 7.7)	4.8*
	Getting an appointment to suit you	77.4 (SD 11.5)	80.2 (SD 1.7)	2.9*
	Getting through to the practice by phone	68.8 (SD 17.3)	72.6 (SD 5.7)	3.8*
	Being able to speak to the GP by phone	63.4 (SD 16.2)	69.6 (SD 5.3)	6.2*
	Waiting time in the waiting room	56.1 (SD 19.9)	61.2 (SD 18.4)	5.0*
	Providing quick services for urgent health problems	83.8 (SD 7.0)	89.2 (SD 7.1)	5.4*
Accessibility in general practice and GP	It is possible to ask for a longer consultation	85.5 (SD 17.9)	91.7 (SD 5.9)	6.1*
	General practitioner is good accessible by phone	71.7 (SD 19.1)	80.4 (SD 13.8)	8.7*
	Patient gets another general practitioner regularly	18.7 (SD 14.7)	18.9 (SD 15.8)	0.2
Procedure for complaints	The practice has an accessible procedure for complaints	52.0 (SD 16.9)	58.9 (SD 11.1)	6.9*
	Mean score Organisation of care (11 items)	72.5 (SD 9.7)	78.2 (SD 8.2)	5.6*

* $p < 0.05$; SD: standard deviation

Discussion

Principal findings

Introduction of a participatory P4P program yielded significant improvements in care delivery. Clinical care indicators, pertaining to both process and outcome, concerning diabetes, COPD, asthma and cardiovascular risk management improved, though only the process indicators were incentivised. The influenza vaccination rate and uptake of cervical cancer screening did not improve significantly. Patients' positive experience with GP's functioning increased significantly on all but one item (n = 16 items), and their positive experience with organisation of care increased significantly on 9 out of 11 items. The room for improvement had a significant impact on actual improvements for both the clinical process and outcome indicators and patient experience.

Practice type had an influence on the change in care processes of COPD and asthma, on the change in cardiovascular risk management outcomes and on the change in patient experience with organisation of care; higher baseline scores were related to lower improvement scores. Practices in large cities improved less than practices in rural area on controlled HbA1c for diabetes patients.

Strengths and weaknesses of our study

Strength of our study is the high follow-up rate of the 60 general practices (92.3%) that participated voluntarily in the P4P program. Nonetheless, the attribution of the improvements to the intervention is complex. The attribution would be much easier if the study design was a randomized controlled trial (RCT). Studies on the effect of P4P often do not meet these criteria.[24] Introducing a control group for a RCT is difficult because collecting data on performance measurements is an intervention strategy on quality improvement in itself and most practices do not like to be in the control arm of such an experiment. This makes our results responsive to other activities in general practice focusing on improving the quality of care of the chosen subjects. However, finding different measures that all changed in the same direction gives us circumstantial evidence that underlines the effect of the program. We did not measure change in practice management as the target users decided to measure practice management only once in 3 years due to the labour intensive data collection. Earlier research showed that the practice management items did not change significantly over a year even if supportive quality improvement projects had been started.[25] So, we may assume that leaving this domain out of the follow-up measurement would not have had large repercussions for the reimbursement.

Comparison with other studies

An accurate comparison with other studies is difficult because performance measures, appraisal and reimbursement differ between P4P programs. The size of

the improvements in chronic care resembles other P4P studies in primary care with stakeholder involvement that showed effect sizes above 10%.[4] Our study shows less improvements with regard to HbA1c testing, probably due to high baseline scores and a limited follow-up period of 1 year.[6,8] We show that baseline measures have a huge impact on the improvements. The lack of improvement in our study in the uptake rate of cancer screening and influenza vaccination is probably also related to the very high baseline in comparison to other studies.[26,27] The comprehensiveness of our indicator set bears a resemblance to the one used in the QOF, but the appraisal and reimbursement are quite different. The QOF bonus is also much higher than the incentive in our intervention. Moreover, in comparison to the QOF, the patient experiences are more related to the payment because the target users decided that patient experiences should become an ample part of quality of care. And as a matter of fact, we did find a bit more improvement on patient experience than the QOF,[28] but this statement should be handled with care because the measurements differed a lot. By incentivising the process indicators, an indirect effect is expected on the reported outcome measures. Our study shows improvements on both clinical process and outcome indicators. This is in line with research of Ryan and Doran (2011), which showed that improving processes of care affect patient outcomes.[29] Although the target users decided not to incentivise the outcome indicators, because a robust judgment on risk adjustment was lacking, the outcome indicators improved as well. This result can contribute to a renewed discussion on the necessity of risk adjustment, which is part of some P4P programs.[30]
Solo and duo practices improved more with regard to COPD and asthma than group practices, which is in line with the literature in which group practices perform better[17,18] and therefore have less room for improvement. When taking into account the underrepresentation of solo practices, this study shows that an even larger effect is possible. Patient's experience with GP's functioning and the organisation of care in the participating practices was very positive, which is in line with other studies.[23,31] The improvement in patient experience due to P4P is hardly studied. In a Cochrane review,[24] only one study could be included in which no effect of P4P on patient experience was shown.[32]

Implications for general practice, policy makers and future research

Though the effectiveness of P4P is inconclusive in the literature, we conclude that a bottom-up developed P4P program might stimulate improvement in clinical care and patient experience with GP's functioning and the organisation of care.
More studies are needed in which the appraisal and reimbursement are based on drivers taken from behavioural economics. Implementation problems as

discussed in QOF research[33,34] such as ceiling effects, reversal effects due to the withdrawal of the payment and the narrow focus on care that is paid for are probably not solved by introducing a participatory P4P program. However, having the target users involved makes it possible to deal with these problems, for instance by renewing the set periodically, which might enlarge the potential effect of P4P. This involvement might even help to reduce unintended consequences like gaming and neglecting the conditions that are not incentivised.[35-37]

References

1
Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S. Does pay-for-performance improve the quality of health care? *Ann Intern Med* 2006; 145: 265-72.

2
Rosenthal MB, Frank RG. What is the empirical basis for paying for quality in health care? *Med Care Res Rev* 2006; 63: 135-57.

3
Frølich A, Talavera JA, Broadhead P, Dudley RA. A behavioral model of clinician responses to incentives to improve quality. *Health Policy* 2007; 80: 179-93.

4
Van Herck P, De Smedt D, Annemans L, Remmen R, Rosenthal MB, Sermeus W. Systematic review: effects, design choices, and context of pay-for-performance in health care. *BMC Health Serv Res* 2010; 10: 247.

5
Amundson G, Solberg LI, Reed M, Martini EM, Carlson R. Paying for quality improvement: compliance with tobacco cessation guidelines. *Jt Comm J Qual Saf* 2003; 29: 59-65.

6
Chung RS, Chernicoff HO, Nakao KA, Nickel RC, Legorreta AP. A quality-driven physician compensation model: four-year follow-up study. *J Healthc Qual* 2003; 25: 31-7.

7
Gilmore AS, Zhao Y, Kang N *et al.* Patient outcomes and evidence-based medicine in a preferred provider organization setting: a six-year evaluation of a physician pay-for-performance program. *Health Serv Res* 2007; 42(6 Pt 1): 2140-59

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

9
Kirschner K, Braspenning J, Jacobs JE, Grol R. Design choices made by target users for a pay-for-performance program in primary care: an action research approach. *BMC Fam Pract* 2012; 13: 25.

10
Mehrotra A, Sorbero ME, Damberg CL. Using the lessons of behavioral economics to design more effective pay-for-performance programs. *Am J Manag Care* 2010; 16: 497-503.

11
Thaler RH. Mental accounting and consumer choice. *Marketing Sci* 1985; 4: 199-214.

12
Heath C, Larrick RP, Wu G. Goals as reference points. *Cogn Psychol* 1999; 38: 79-109.

13
Loewenstein G, Prelec D. Anomalies in intertemporal choice: evidence and an interpretation. *Q J Econ* 1992; 107: 573-97.

14
Thaler RH. Mental accounting matters. *J Behav Decis Making* 1999; 12: 183-206.

15
Chung S, Palaniappan LP, Trujillo LM, Rubin HR, Luft HS. Effect of physician-specific pay-for-performance incentives in a large group practice. *Am J Manag Care* 2010; 16: e35-42.

16
Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979; 47: 263-92.

17
Ashworth M, Armstrong D. The relationship between general practice characteristics and quality of care: a national survey of quality indicators used in the UK Quality and Outcomes Framework, 2004-5. *BMC Fam Pract* 2006; 7: 68.

18
Ashworth M, Schofield P, Seed P, Durbaba S, Kordowicz M, Jones R. Identifying poorly performing general practices in England: a longitudinal study using data from the quality and outcomes framework. *J Health Serv Res Policy* 2011; 16: 21-7.

19
Campbell SM, Hann M, Hacker J *et al.* Identifying predictors of high quality care in English general practice: observational study. *BMJ* 2001; 323: 784-7.

20
Van den Hombergh P. *Practice Visits. Assessing and Improving Management in General Practice*. Nijmegen: Catholic University of Nijmegen, 1998.

21
Grol R, Dautzenberg M, Brinkmann H, eds. *Quality Management in Primary Care. European Practice Assessment.* Gutersloh: Verlag Bertelsmann Stiftung, 2004.

22
Grol R, Wensing M, Mainz J *et al.* Patients' priorities with respect to general practice care: an international comparison. European Task Force on Patient Evaluations of General Practice (EUROPEP). *Fam Pract* 1999; 16: 4-11.

23
Grol R, Wensing M, Mainz J *et al.*; European Task Force on Patient Evaluations of General Practice Care (EUROPEP). Patients in Europe evaluate general practice

care: an international comparison. *Br J Gen Pract* 2000; 50: 882-7.

24
Scott A, Sivey P, Ait Ouakrim D *et al.* The effect of financial incentives on the quality of health care provided by primary care physicians. *Cochrane Database Syst Rev* 2011; 9: CD008451.

25
Engels Y, van den Hombergh P, Mokkink H, van den Hoogen H, van den Bosch W, Grol R. The effects of a team-based continuous quality improvement inter-vention on the management of primary care: a randomised controlled trial. *Br J Gen Pract* 2006; 56: 781-7.

26
Loerbroks A, Stock C, Bosch JA, Litaker DG, Apfelbacher CJ. Influenza vaccination coverage among high-risk groups in 11 European countries. *Eur J Public Health* 2012; 22: 562-8.

27
Van Ballegooijen M, van den Akker-van Marle E, Patnick J *et al.* Overview of important cervical cancer screening process values in European Union (EU) countries, and tentative predictions of the corresponding effectiveness and cost-effectiveness. *Eur J Cancer* 2000; 36: 2177-88.

28
Campbell SM, Kontopantelis E, Reeves D, Valderas JM, Gaehl E, Small N *et al.* Changes in patient experiences of primary care during health service reforms in England between 2003 and 2007. *Ann Fam Med* 2010; 8: 499-506.

29
Ryan AM, Doran T. The effect of improving processes of care on patient outcomes: evidence from the United Kingdom's quality and outcomes framework. *Med Care* 2012; 50: 191-9.

30
Zaslavsky AM, Hochheimer JN, Schneider EC *et al.* Impact of sociodemographic case mix on the HEDIS measures of health plan quality. *Med Care* 2000; 38: 981-92

31
Allan J, Schattner P, Stocks N, Ramsay E. Does patient satisfaction of general practice change over a decade? *BMC Fam Pract* 2009; 10: 13.

32
Gosden T, Sibbald B, Williams J, Petchey R, Leese B. Paying doctors by salary: a controlled study of general practitioner behavior in England. *Health Policy* 2003; 64: 415-23.

33
Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M. Effects of pay for performance on the quality of primary care in England. *N Engl J Med* 2009; 361: 368-78.

34
Roland M, Campbell S, Bailey N, Whalley D, Sibbald B. Financial incentives to improve the quality of primary care in the UK: predicting the consequences of change. *Prim Health Care Res Dev* 2006; 7: 18-26.

35
Campbell S, Reeves D, Kontopantelis E, Middleton E, Sibbald B, Roland M. Quality of primary care in England with the introduction of pay for performance. *N Engl J Med* 2007; 357: 181-90.

36
Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of patients from pay-for-performance targets by English physicians. *N Engl J Med* 2008; 359: 274-84.

37
Conrad DA, Perry L. Quality-based financial incentives in health care: can we improve quality by paying for it? *Annu Rev Public Health* 2009; 30: 357-71.

38
Hingstman L, Kenens RJ. *Cijfers uit de registratie van huisartsen. Peiling 2008*. [Figures from the general practitioners, 2008]. Utrecht: NIVEL, 2008.

Chapter 6

Experiences of general practices with a participatory pay-for-performance program: a qualitative study in primary care

Kirsten Kirschner

Jozé Braspenning

JE Annelies Jacobs

Richard Grol

Australian Journal of Primary Health 2013;19(2):102-106

Abstract

The involvement of target users in the design choices of a pay-for-performance program may enhance its impact, but little is known about the views of participants in these programs. To explore general practices' experiences with pay-for-performance in primary care we conducted a qualitative study in general practices in the Netherlands. Thirty out of 65 general practices participating in a pay-for-performance program, stratified for bonus, were invited for a semistructured interview on feasibility, feedback and the bonus, spending of the bonus, unintended consequences, and future developments. Content analysis was used to process the resulting transcripts. We included 29 practices. The feasibility of the pay-for-performance program was questioned due to the substantial time investment. The feedback on clinical care, practice management and patient experience was mostly discussed in the team, and used for improvement plans, but was also qualified as annoying for one GP and for another GP it brought feelings of insecurity. Most practices considered the bonus a stimulus to improve quality of care, in addition to compensation for their effort and time invested. Distinctive performance features were not displayed, for instance, on a website. The bonus was mainly spent on new equipment or team building. Practices referred to gaming and focusing on those aspects that were incentivised ('tunnel vision') as unintended consequences. Future developments should be directed to absolute thresholds, new indicators to keep the process going, and an independent audit. Linking a part of the bonus to innovation was also suggested. The participants thought the pay-for-performance program was a labour intensive positive breakthrough to stimulate quality improvement, but warned of unintended consequences of the program and the sustainability of the indicator set.

Introduction

In recent decades health policy makers in various countries have introduced pay-for-performance (P4P) programs with the aim to enhance the quality of care. However, the evidence for their effectiveness is still lacking.[1] Most studies have deficiencies in their design and the payment schemes (content and amount) are very heterogeneous.[1,2] A review showed that involving target users could increase a P4P program's effectiveness.[3] We conducted a project in the Netherlands in which we involved GPs in the development of a P4P program for primary care. The resulting program included indicators for clinical care, practice management and patient experience.[4] Its key elements are presented in Box 1.

Box 1 Design choices of the participatory P4P program

Performance measurement

- ***Data collection for clinical care, practice management and patient experience***
  - Clinical care: diabetes (n = 9 indicators), COPD (n = 5 indicators), asthma (n = 4 indicators), cardiovascular risk management (n = 8 indicators), influenza vaccination (n = 2 indicators), cervical cancer screening (n = 1 indicator)
  - Practice management (n = 4 indicators): infrastructure (n = 7 items), team (n = 8 items), information (n = 3 items), quality and safety (n = 4 items)
  - Patient experience (n = 2 indicators): experience with general practitioner (n = 16 items) and organisation of care (n = 11 items)

Appraisal

- Separate appraisal of clinical care, practice management and patient experience, weighting the domains as 2:1:1
- A benchmark with relative standards was set at the 25th percentile of group performance
- For the appraisal a series of tiered thresholds was used (7 levels)
- Practices received short-term feedback (4 months after data collection)
- Both quality level and the improvement of performance were valued, with these levels weighted as 3:1

Reimbursement

- A bonus of 5% to 10% of the practice income, not linked to the usual reimbursement
- Bonus was paid in money, and not in goods or services
- Bonus to spend freely

The clinical indicators were related to chronic care, prevention and prescribing and are largely similar to those initiated in the USA[5-10], Australia[11], New Zealand[12] and the UK[13]. The appraisal and reimbursement features of our P4P program were tailored to the target group and differ from other programs on the performance thresholds (relative), the size of the bonus, and incentivising both quality level and improvement of performance. The question was how the target users evaluated the actual use of the P4P program that they helped to design, and if it met their expectations. Studies about the experiences of target users with P4P programs are scarce. Physicians participating in P4P programs in Massachusetts and California showed positive attitudes towards P4P, but were ambivalent about specific features of these programs.[14] General practices participating in the

Australian Practice Incentives Program (PIP) were asked about their views on PIP's contribution to quality of care and improved access. Their views were mixed, with 27% of providers responding that PIP gave significant benefit to their practice, 36% responding that there was medium benefit, and 27% responding that the benefit was minor.[15] Campbell et al.[16] interviewed GPs and nurses about their views on changes in health care as a result of the Quality and Outcomes Framework.[16] The respondents believed that the financial incentives had been sufficient to change behaviour and to achieve targets, but they also mentioned some unintended consequences, such as a decline in personal continuity of care. Furthermore, the interviewees worried about an ongoing culture of performance monitoring in the UK. Little is known about how target users who are involved in designing a P4P program evaluate that program after implementation. In our study we asked representatives of participating practices to provide their experiences.

Methods

Study design and study population

Sixty-five general practices participated voluntarily in the development and use of a P4P program, and received a financial incentive afterwards. We performed a qualitative study based on individual semistructured interviews with representatives of these general practices. We randomly invited 30 of the 65 general practices for the interviews, stratified for the bonus they received (low versus high).

Interview

The interview guide contained questions about the feasibility of the data collection, behavioural change concerning the feedback on the indicators and the bonus, spending of the bonus, possible unintended consequences of the program and future developments. The interview was semistructured, with open questions, but the interviewers were prepared and trained to address the different topics in more detail. The interviews were performed by three trained professionals who collected data from April to June 2008.

Data analysis

The interviews were audio recorded and transcribed verbatim. The topics in the interview guide were used as a coding frame. Codes were linked to the data using the software package ATLAS ti.[17] The trained professionals performed one interview to test its length and to make sure the information drawn from the interview was sufficient. After three interviews the interview guide was reassessed, but adjustments were not deemed necessary. Quotes from the general practices (P) are added to the description of the results, with the practice

number accompanied by an 'a' for practices that received a low bonus and a 'b' for practices with a high bonus.

Results

Study population

An appointment was scheduled for each of 30 general practices. We succeeded in interviewing representatives of 29 general practices, 14 of which received a low bonus and 15 a high bonus. In 28 practices the GP was interviewed, sometimes together with another GP (n = 4) or a practice nurse (n = 2). In one practice the interview was held with a practice nurse.

Feasibility

For most of the practices the data collection for the P4P program meant a substantial time investment, especially for the measures of clinical care. General practices experienced difficulties with extracting data from their electronic medical records (EMR) caused by a lack of extraction software for the EMR and their deficiency in the uniformity of registration.

> *We experienced many problems with extracting the data from the electronic medical records, it took a lot of time. We developed many plans to improve recording. (P21a)*

> *At baseline, putting together the figures of our practice, we spent at least 4 weekends between the two of us. (P1b)*

If practices could not extract the data from their EMR automatically they had to take a random sample of 40 patients for each chronic condition, which took even more time.

> *We had to do everything manually. I think the assistants used a hundred hours on top of their normal working hours for this. The time investment was disappointing. (P5a)*

Although practices complained about the time investment, we noticed that some practices provided complete data manually on all patients instead of a sample.

Behavioural change

Feedback showed general practices their performance for clinical care, practice management and patient experience. One GP reported:

> *I thought I was doing well, but now I got more insight into what really happens. (P1b)*

The feedback was generally seen as confirmation of the normal routines. GPs concluded that they were performing reasonably well. Some parts of the feedback were unexpected and at times difficult to deal with. This usually resulted in plans for improvement. One GP told us he became more insecure, because the feedback of patients was negative. Another GP thought it was annoying that some things were not optimal in his practice. GPs discussed the feedback within their teams, which resulted in good team climates. Discussing the feedback was seen as good closure after an intensive period of data collection in which the whole team participated.

The bonuses varied from 900 to 40.000 Euros. Not all GPs knew whether they had received the bonus. The perceived impact of the bonus differed between GPs. Some GPs stated that the bonus had had no influence on performance, but most GPs saw the bonus as a stimulus to improve quality and to perform better the next year. Most GPs felt appreciated by the bonus. 'We always need to go faster and do more' one explained, 'We have to deliver quality but it should not cost more. So, the incentive is a nice stimulus to do this all' (P17b).

The practices reported performance information at an aggregated level, that is separate scores on clinical care, practice management and patient experiences. Although public reporting was not part of the program we asked GPs to reflect on being transparent to other parties. Most GPs discussed their performance scores in detail with their colleagues within the practice, but very few practices made their distinctive features public outside the practice. About being transparent to colleagues outside the practice, one GP said:

> *Well, that is a sensitive topic. When it is about whether there is enough privacy in our practice, everyone should know. When it is about how many consultations the GP has per week and about waiting time in the waiting room, I do not think others should know. That is non-public. (P21a)*

About half of the GPs stated that they do not want to give detailed information on their performance to insurers, fearing that the information might be used for sanctions or penalties. Others were fine with providing such information. A major concern about transparency to patients was related to whether patients would understand the information presented:

> *Well, I am not sure whether it is information for the newspaper. When a patient wants to see it, he or she has to be able to judge it on the content. (P16b)*

GPs were convinced that patients should be informed about quality of care, but how this information should be presented was not clear.

Spending of the bonus

The practices could spend the bonus in any way they liked, because there were no restrictions. The bonus was seen as a reward, a gift. Most practices used the bonus as extra practice income and bought equipment for the practice. Some practices used the bonus for a dinner or a weekend trip for the whole team. One GP bought a coffee machine for the patients in the waiting room. In some practices the incentive ended up with an individual GP and was used for private matters.

Unintended consequences of P4P

One of the unintended consequences mentioned was 'gaming' the system. Gaming can be caused by a fear of lost reputation, but also by financial reasons. Many GPs imagined other practices gaming the data by saying:

> *Oh, that will happen for sure. (P5a)*

> *For sure, yes. I am convinced. (P14a)*

But when asked if they gamed their own data, most of them said resolutely 'no'. One GP admitted that:

> *We manipulated those data of which we thought that they were not correct at first. You try to get the most out of it. (P18b)*

Most participants thought data should also be verified by an external party.

> *If there are any consequences, data should definitely be checked by an external party. When there is a financial reward I think the figures should be correct. (P12b)*

Others thought it a matter of trust and the responsibility of the GPs to make sure their figures were correct.

> *You are only getting yourself into trouble if you do not provide data that is correct. (P21a)*

Another unintended consequence mentioned was developing 'tunnel vision', which meant that GPs focussed on those aspects of care that were incentivised. For example:

> *If you reward me for rinsing ears, I will do that more often even though it is not useful. (P22a)*

We looked up which patient should have a risk profile and checked if they answered it. If they did not, we gave them a call and asked them to come over for a venipuncture because that was missing for example. Eventually, we got a fully completed risk profile. You could label it as cheating. We had a goal on which we wanted to perform well and therefore we invited those patients. When there was no goal we would not have invited that patient. (P4b)

Another GP said:

If a diabetic sits in front of me, I now think of prescribing a statin. (P19a)

Respondents thought that focusing on those aspects of care that were incentivised might be done at the expense of other aspects that might be more important, like personal attention.

Future developments

In deciding on the P4P design choices, the majority of respondents preferred relative performance thresholds, i.e. a relative benchmark retrieved from participants' scores. In the interviews, however, we saw a switch towards a preference for absolute thresholds. Some GPs thought the relative thresholds were set at a high level. Comparison with absolute thresholds might be more honest, assuming that these will not be raised when group performance improves. Nevertheless, the interviewees expected a lot of debate in setting absolute thresholds, for example on how to determine the thresholds and who to involve in this process. According to the respondents the thresholds should definitely not be set by the government or by health insurance companies and the professionals should be involved in that process.

The professional group should determine the thresholds. The thresholds can be raised gradually to see if everyone can reach them. If that is the case, the thresholds can be raised again. In this way quality improves while everyone can meet the thresholds. (P11b)

GPs were positive about the further development of P4P. P4P was seen as a breakthrough in the Netherlands. Most GPs thought that both quality level and improvement of performance should be rewarded and that there should be an incentive to continue improving patient care. Linking the incentive to innovation was also suggested. Some GPs reported that practices that did not perform well should be penalised in future. Transparency was a positive thing to most GPs, because it can instigate discussion on the quality of care given. Some concerns were mentioned, for instance regarding the focus of the incentive. According to

the GPs the incentive should focus on a wide variety of aspects of care to avoid 'tunnel vision'. One GP was concerned about the UK P4P program in which GPs are paid for asking a patient if they smoke, but not for motivating them to stop smoking.

Discussion

Summary of main findings

One year after the introduction of the P4P program we gathered feedback from 29 general practice representatives who participated in designing the program. They considered the feedback and the bonus as stimuli to improve quality and perform better the next year, though the data collection proved to be a huge time investment. The GPs felt appreciated by the bonus and invested the bonus mostly in their practice. Very few practices made their figures public outside the practice. The GPs were positive about further development of the P4P program, but unintended consequences like gaming and developing 'tunnel vision' should be taken into account. GPs have little confidence in the health insurance companies and fear that their information will be used for sanctions or penalties.

Strengths and limitations of the study

The strength of this study lies in its qualitative approach, which allowed us to get a clear picture of GPs' experiences with the participatory P4P program. Our sampling method was designed to collect the views of those GPs who had actually worked with the P4P program, but as we assumed that the amount of bonus received might influence the views of the participating GPs, we included an equal number of high and low bonuses paid to GPs. There were no differences in views between the two groups.

Limitations arose because time restrictions hindered the discussion of each topic in detail. On more general topics, like the future of P4P, there was some discussion but concrete ideas were not brought up. Another limitation was the lack of a comparative group that was not involved in designing the program, and that did not participate voluntarily in the program. It would be of interest to learn about the experiences of such a group. It can be argued that this group would probably be less enthusiastic. However, based on our results we could also argue that an implementation strategy would benefit the involvement of target users in designing the program and lead to positive experiences.

Comparison with existing literature

To date no study has published the experiences of GPs who were actually involved in the development of a P4P program. The involvement of the target users resulted in a high awareness of the P4P program and the target contents. This may seem obvious, yet we found major differences in the extent to which

physicians in other countries were aware of the targets of a P4P program.[18-20] The positive attitude of the target users towards P4P in our study is in line with the results in the study of Young et al.[14] The criticisms of our program in terms of the administrative workload, relative performance thresholds and unintended consequences were also found in the study of Campbell et al.[16] and in a review of the administrative burden of the PIP.[22] On the whole, the participating GPs were positive about the P4P program. However, involvement of the target users in the development process did not ensure that there were no concerns about the future of P4P.

Implications for future research or clinical practice

The experiences of general practices with P4P have taught us some lessons. There is a tension between incentivising improvement of performance on specific indicators and a P4P program with a broad scope of aspects on the one hand and less administrative workload and securities (less variety in aspects) in the long run on the other hand. It is important to have a balanced P4P program to avoid 'tunnel vision' and to maintain and improve quality of care in as many aspects as possible. The administrative workload can be reduced by not measuring the domain of practice management every year but only once every 3-5 years. Furthermore, the problem of time investment in data extraction from medical records can be resolved by introducing suitable equipment, as was done in the UK.[13] For general practices to have more confidence in the health insurance companies longer-term contracts might be a solution. Currently, agreements with the health insurance companies are still made as 1-year contracts. This issue might be specific to the Dutch situation, but having confidence in payers' policies will be an issue in other countries as well.

According to the GPs, absolute thresholds are preferred.[22,23] They have the advantage that there is no uncertainty whether a threshold has been met[24] and GPs know exactly which thresholds to meet and to which aspects of care to improve. A disadvantage of absolute thresholds is that the motivational effects are uncertain.[25] However, this argument can be made for relative thresholds as well, that is being at the very top or bottom discourages quality improvement. So, the difference between relative and absolute thresholds is mainly in how clear the goals are.

Conclusion

The participants thought the P4P program was a labour intensive positive breakthrough to stimulate quality improvement, but warned of unintended consequences of the program and the sustainability of the indicator set of the P4P program.

References

1
Scott A, Sivey P, Ait OD, Willenberg L, Naccarella L, Furler J, et al. The effect of financial incentives on the quality of health care provided by primary care physicians. *Cochrane Database Syst Rev* 2011;9:CD008451.

2
Mehrotra A, Sorbero ME, Damberg CL. Using the lessons of behavioral economics to design more effective pay-for-performance programs. *Am J Manag Care* 2010;16(7):497-503.

3
Van Herck P, De SD, Annemans L, Remmen R, Rosenthal MB, Sermeus W. Systematic review: effects, design choices, and context of pay-for-performance in health care. *BMC Health Serv Res* 2010;10:247.

4
Kirschner K, Braspenning J, Jacobs JA, Grol R. Design choices made by target users for a pay-for-performance program in primary care: an action research approach. *BMC Fam Pract* 2012;13(1):25.

5
Dudley RA, Luft HS. Managed care in transition. *N Engl J Med* 2001;344 (14):1087-92.

6
Epstein AM, Lee TH, Hamel MB. Paying physicians for high-quality care. *N Engl J Med* 2004;350(4): 406-10.

7
Galvin R, Milstein A. Large employers' new strategies in health care. *N Engl J Med* 2002;347(12): 939-42.

8
Wetzel S, Galvin R, Buck CR, Jr., Cubbin J, Bradley B, Taylor B, et al. Taking a giant leap forward in promoting quality. *Health Aff (Millwood)* 2000;19(2):275-76.

9
Baker G. *Pay for performance incentive programs in healthcare: market dynamics and business process*. San Fransisco, CA, 2004.

10
Excellence Bt. *Rewarding quality across the healthcare system*. Bridges to Excellence, 2009.

11
Australia M. *Practice Incentives Program (PIP).* Medicare Australia, 2009.

12
Buetow S. Pay-for-performance in New Zealand primary health care. *J Health Organ Manag* 2008;22(1): 36-47.

13
Roland M. Linking physicians' pay to the quality of care: a major experiment in the United kingdom. *N Engl J Med* 2004;351(14):1448-54.

14
Young GJ, Meterko M, White B, Bokhour BG, Sautter KM, Berlowitz D, et al. Physician attitudes toward pay-for-quality programs: perspectives from the front line. *Med Care Res Rev* 2007;64(3):331-43.

15
Office ANA. *Practice Incentives Program.* Department of Health and Ageing. Medicare Australia. Audit Report No.5 2010-11, 2010.

16
Campbell SM, McDonald R, Lester H. The experience of pay for performance in English family practice: a qualitative study. *Ann Fam Med* 2008;6(3):228-34.

17
Atlas.ti. *Qualitative data analysis*, 2011.

18
McDonald R, White J, Marmor TR. Paying for performance in primary medical care: learning about and learning from 'success' and 'failure' in England and California. *J Health Polit Policy Law* 2009;34(5):747-76.

19
Teleki SS, Damberg CL, Pham C, Berry SH. Will financial incentives stimulate quality improvement? Reactions from frontline physicians. *Am J Med Qual* 2006;21(6):367-74.

20
McDonald R, Roland M. Pay for performance in primary care in England and California: comparison of unintended consequences. *Ann Fam Med* 2009;7(2):121-27.

21
Association AM. *AMA Submission to the Productivity Commission study on administrative and compliance costs associated with Commonwealth programs that impact specifically on general practice*, 2002.

22
Dudley RA, Frolich A, Robinowitz DL, Talavera JA, Broadhead P, Luft HS. Strategies to support quality-based purchasing: a review of the evidence. *Technical Review* No.10 ed. Rockville, MD, 2004.

23
Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S. Does pay-for-performance improve the quality of health care? *Ann Intern Med* 2006;145(4):265-72.

24
Hahn J. *Pay-for-Performance in Health Care*. Washington, DC, 2006.

25
Young GJ, Meterko M, Beckman H, Baker E, White B, Sautter KM, et al. Effects of paying physicians based on their relative performance for quality. *J Gen Intern Med* 2007;22(6):872-76.

Chapter 7

General discussion

The main objective of this thesis was to study the effects of a participatory pay-for-performance (P4P) program on the improvement of the quality of care. This chapter presents the main results of the different projects, their strengths and weaknesses, and the key implications for practice and research.

Main findings

Effect of practice-based improvement plans on accessibility and availability in general practice (Chapter 2)

- Feedback and best practice examples can stimulate improvement in the organisation of general practice care, but their overall effects on change were modest.

The design choices of P4P: performance measures, appraisal and reimbursement (Chapter 3)

- A P4P program can be designed by involving the target groups of practitioners and the payers (health insurance companies) at the same time. The target users discussed performance measures, appraisal and reimbursement. The final design choices are described in Box 1.
- The resulting design resembled the P4P programs in other countries, but seemed to be more balanced with regard to the position of patient experiences in the program.

Box 1 Design choices of the participatory P4P program

Performance measurement

- ***Data collection for clinical care, practice management and patient experience***
  - Clinical care: diabetes (n = 9 indicators), COPD (n = 5 indicators), asthma (n = 4 indicators), cardiovascular risk management (n = 8 indicators), influenza vaccination (n = 2 indicators), cervical cancer screening (n = 1 indicator)
  - Practice management (n = 4 indicators): infrastructure (n = 7 items), team (n = 8 items), information (n = 3 items), quality and safety (n = 4 items)
  - Patient experience (n = 2 indicators): experience with general practitioner (n = 16 items) and organisation of care (n = 11 items)

Appraisal

- Separate appraisal of clinical care, practice management and patient experience, weighting the domains as 2:1:1
- A benchmark with relative standards was set at the 25th percentile of group performance
- For the appraisal a series of tiered thresholds was used (7 levels)
- Practices received short-term feedback (4 months after data collection)
- Both quality level and the improvement of performance were valued, with these levels weighted as 3:1

Reimbursement

- A bonus of 5% to 10% of the practice income, not linked to the usual reimbursement
- Bonus was paid in money, and not in goods or services
- Bonus to spend freely

Validity and reliability of the indicator set (Chapter 4)

- The evaluation of the clinical measures showed that:
  - A broad set of aspects of clinical care were measured
  - Each topic was described by a coherent set of indicators
  - To allow a 10% error margin at least 96 patients should be assessed
  - A reliable benchmark (5% error allowed) should be based on at least 233 practices.
- The measures for practice management (Visitation Instrument Practice Management (VIP) and patient experience (EUROPEP instrument) were validated in previous research.

Effect of the P4P program: assessment of indicators and experiences (Chapter 5 and 6)

- The performance of clinical care and patients' experience of GP functioning and the organisation of care significantly improved following the introduction of the P4P program. Significant improvements were observed for the process indicators for diabetes (+10.4%), COPD (+8.1%), asthma (+11.5%) and cardio-vascular risk management (+7.9%). Patients' experience of GP functioning statistically improved by 5.9% and patients' experience of the organisation of care by 5.6%.
- Five out of seven outcome indicators showed significant improvements, ranging from +5.9% (blood pressure diabetes) to +14.7% (blood pressure CVRM).
- General practitioners considered the feedback and the bonus to be a stimulus for improving quality and performance in the next year, although the data collection involved a substantial time investment. The participants were positive about the further development of the P4P program, but unintended consequences like manipulating behaviour at work in response to the program (gaming) and too narrow a focus on program objectives ('tunnel vision') should also be taken into account.

Quality improvement: intrinsic versus extrinsic motivation

To improve the quality of care in general practice, it is necessary to change the behaviour of care providers. Different interventions are known to stimulate quality improvement. These interventions can be divided into those that adopt educative, informative and facilitative methods aimed at enhancing the intrinsic motivation for quality improvement of the healthcare providers, and those that use methods that are more controlling where the aim is to raise the extrinsic motivation of healthcare providers.[1] Giving professionals feedback about their performance is an example of stimulating the intrinsic motivation, whereas

financial or material rewards, promoting competition and the public reporting of performance are common methods for stimulating extrinsic motivation. Audit and feedback are widely used as strategies to improve professional practice, either on their own or as components of multifaceted quality improvement interventions.[2] The Cochrane review by Ivers et al. showed that audit and feedback generally lead to small but potentially important improvements of about 4.3% in the targetted behaviour. Feedback may be more effective when it includes both explicit targets and an action plan.[2] To increase the effects of audit and feedback, benchmarks and information about best practice can be added.[3] In our first study we stimulated the accessibility to general practice by giving the practices feedback accompanied by benchmarking against their peers and information about best practice. This approach led to practice-based improvement plans mostly concerning the information service, waiting times in the waiting room and phone accessibility. These plans were experienced as a stimulus for improving accessibility and availability, although the overall impact of the interventions on the patient experience was small.

Disappointed by these results of stimulating intrinsic motivation on quality improvement, interventions may focus on extrinsic motivation factors such as financial incentives. In this new generation of interventions financial incentives are linked to audit and feedback. Adding a financial incentive to the childhood immunization coverage rate in a setting where there is audit and feedback led to a significant improvement (from 29% to 54% coverage) in the bonus group after eight months.[4] However, no effect was found in a study on implementing guidelines for cancer screening.[5] Overall, the effectiveness of interventions that are focused on raising extrinsic motivation for quality improvement, such as pay-for-performance (P4P) programs, is inconclusive.[6,7] Scott et al. (2011) found seven studies showing the modest and variable effects of P4P on the quality of healthcare.[8] There are reviews that give some suggestions about the possible elements of a successful P4P. Petersen et al. (2006) and Van Herck et al. (2010) suggest the use of combined incentives for both overall improvement and the achievement of a threshold, and that use should be made of process and outcome indicators as target measures.[6,9] Furthermore, P4P targets should be based on defined baseline data and the room for improvement, the implementation of a uniform P4P design across payers, the distribution of incentives at the individual and/or team level, and communication of the program widely and directly during development, implementation and evaluation.[9] The successes depend on the design of the P4P program in which decisions need to be made about performance measures, the appraisal (unit of assessment, performance standards) and the reimbursement (size of the bonus).[10]

Stakeholder involvement can play a promising role in the designing of an optimal P4P program.[9] We assumed that involving GPs in our program would improve

intrinsic motivation, and that adding an incentive to performance would stimulate extrinsic motivation. This hypothesis seems to be confirmed to a large extent. Our P4P program resulted in significant improvements of an average of 6% in the patient experience and 10% in clinical care. Compared to other programs aimed at either intrinsic or extrinsic motivation, the results of our P4P experiment resulted in greater changes. The review by Van Herck et al. showed that P4P programs resulted in about 5% improvement on average, but with a wide variation in the amount of change depending on the measure and the program.[9] They also found that studies with stakeholder involvement had more positive effects (above 10% effect size) than those that did not involve stakeholders, which is in line with our results.[9]

Stakeholder involvement in designing a P4P program

A unique experiment was conducted in 65 general practices in the south of the Netherlands. In this experiment financial incentives were thus linked to a set of performance indicators. The P4P program was designed systematically with the involvement of the stakeholders (health insurance companies and GPs).[11] In panel discussions we discussed the content of the P4P program, more specifically the performance measures, appraisal and reimbursement. An important question to ask is: Is the design of a P4P program actually affected by stakeholder involvement?

Performance measures

The involvement of stakeholders in the design of the P4P program had only a partial effect on the type of quality indicators chosen. Based mainly on the national accreditation program for primary care and the Quality and Outcomes Framework (QOF) in the UK, the panel was quite clear that three domains of performance measures should be distinguished: clinical care, practice management and patient experience.[12,13] Other P4P programs have featured comparable domains.

In the QOF, income is linked to performance based on 76 clinical quality indicators and a further 70 indicators relating to the organisation of care and patient experience. The clinical indicators mainly relate to processes – for example measuring disease parameters and giving treatment – with only 10 of the 76 original clinical indicators relating to outcomes.[12] The Practice Incentive Program (PIP) in Australia includes incentives for 11 aspects of care including the quality of prescribing, diabetes, asthma, cervical cancer, indigenous health, e-health, after-hours care, teaching, rural loading, aged care access, and a financial incentive aimed at ensuring access to surgical, anaesthetic, and obstetric services in rural areas.[14] The Integrated Healthcare Association (IHA) P4P program in the United States contains four measurement domains (measurement year 2012): clinical

quality (prevention, cardiovascular, diabetes, maternity, musculoskeletal, and respiratory conditions), patient experience, meaningful use of health IT and appropriate resource use.[15]

In these programs clinical care measures focus on chronic care and prevention, as in our program. Not all programs involve patients' experiences. In Australia, for example, patients' experiences are no part of the PIP. With respect to patient centred care it is important to know patients' experiences. This makes it possible to actually listen to your patient and act accordingly. Furthermore, our program seems to be more balanced than other programs with regard to the weighting of clinical care, practice management and patient experiences (2:1:1) in the program.[12,14,16,17] For example, in the QOF the weighting of patient experiences is much lower than in our program.

The involvement of stakeholders led to the inclusion of a smaller number of indicators in our program than the UK's program, but to a greater number than in Australia and the US. This result possibly reflects the struggle to obtain meaningful information in a feasible manner. Obtaining meaningful information appeals to intrinsic motivation, but it can be restricted by the data collection, which needs to be feasible and not too labour intensive. The panel suggested that the performance measures could be rotated every couple of years to include indicators that cover a wide range of aspects of care while still keeping the data collection feasible.

Appraisal and reimbursement

The involvement of the stakeholders in the design of the P4P program resulted in three main differences with other programs, which are: (1) the use of relative performance standards, (2) rewarding both quality level and performance improvement, and (3) the size of the bonus.

The thresholds that were set were 'relative', based on the 25th percentile of group performance, which is in contrast to other P4P programs that mainly determine their incentives on the basis of absolute performance standards.[6,8,14,16-19] In the UK minimum thresholds were set at an indicator compliance score of 40%, although a few at 25%, and maximum thresholds were set at an indicator score of 90%.[17] Practices earned points based on the proportion of eligible patients for whom the quality targets were achieved.[12] Relative thresholds can provoke uncertainty and complexity, which can negatively influence the effectiveness of the P4P program.[20] In our panel discussion a slight majority preferred the relative threshold applied in a way that could prevent low achievers from dropping out because they thought that the benchmark outranged their performance capabilities. In the interviews with the GPs that took place after awarding the bonus, the majority seemed to prefer absolute thresholds, which were perceived as more honest, because they would not be raised when group

performance improves. Lessons from behavioural economics also tend to favour the formulation of absolute thresholds in P4P programs to enhance their effectiveness.[20] It is stated that a program with absolute performance standards is more secure for the participants, because the level of performance required to earn the incentive is known.

Our stakeholders chose to incentivise both the quality level and the level of performance improvement, while most other programs reward only good performance but not improvement.[21] The Integrated Healthcare Association (IHA) P4P program also rewards improvement based on changes in the performance between two consecutive years.[15] A combination of both quality level and performance improvement was considered to be more likely to gain acceptance. Furthermore, this mix of reward criteria stimulates improved practices at the bottom as well as at the top. It reflects the focus on intrinsic motivation very well. It was considered that all participants should have something to strive for, otherwise the interest in the program would fade away and quality improvement would be hard to achieve.

The discussion with the target users about the size of the bonus resulted in a bonus of on average 5% to 10% of practice income. The size of the bonus should not be too small as this may limit its effects, nor too high because of unintended consequences like gaming. It was agreed that the proposed size of the bonus should be in line with the bonuses paid in trade and industry. GPs and insurers could easily reach a consensus about a bonus of, on average, 5% to 10% of practice income once the argument for this was put on the table. Our bonus was much smaller than the incentive in the UK which makes up approximately 25% of GPs' income.[22] In other programs bonuses vary between US$ 2 per patient and US$ 10000 per practice.[23] Our practices could earn between € 0 and € 6890 (about US$ 8800) per 1000 patients.[24]

Stakeholder involvement and P4P: critical remarks

Stakeholder involvement in the design of a P4P program that combines intrinsic and extrinsic motivation seems to have influenced the effectiveness of the P4P program in a positive way. Nevertheless, critical remarks can be made as well. First of all, questions can be raised about the selection of the stakeholders. Chronic diseases played a prominent role, therefore it might have been worthwhile to invite (and assess) other caregivers such as the practice nurse, the dietician, the physical therapist and the pharmacist. The patients were invited to give written responses during the determination of the set of indicators in the context of the Dutch National GP Accreditation program, but they were not invited to the panel discussion. Perhaps, their contribution could have helped avoid the omission of inviting other caregivers. Two of the five largest health insurance companies took part in the design of the P4P program. During the

panel discussions the insurers were a little hesitant, especially when the clinical indicators were discussed. It might have been worthwhile involving other health insurers as well, and to involve insurers better in some decisions, especially to enhance the national roll out. Concerning national roll out, the sustainability and wider implementation of the program largely failed: doubts were raised about the long-term effects, the side-effects as well as the registration burden. The P4P program was seen as an interesting project but there was a lack of drive to implement it any further. The various stakeholders did not want to take the initiative, but were waiting for a national initiative by the government.

We interviewed GPs about their experience with the P4P program and the decisions made after they participated in it. The involvement of target users in the developmental process proved to be no guarantee that there was no criticism of the program. Participating in a P4P program does not necessarily mean that participants are fully aware of the program. Major differences were found in the extent to which physicians in other countries were aware of the (targets of a) P4P program, which is an issue that can affect the impact of the program.[25-27] At least their involvement in designing our program has led to a higher awareness of its existence among the participating practices. So far no studies have been published about experiences of GPs who were actually involved in the development of a P4P program. However, studies about the experiences of participation in a P4P program are available. Physicians in the study of Young et al. agreed that physicians should be financially rewarded for providing a higher quality of care: P4P could improve quality and it is more effective than peer recognition alone.[28] Participants of the QOF believed the financial incentives had been sufficient to change behaviour and to enable targets to be achieved.[29]

Criticism of our P4P program and its unintended consequences mostly concerned gaming the system, developing 'tunnel vision', the administrative workload and the relationship between provider and payer. Gaming can be caused by the fear of losing one's reputation, but also by financial considerations. The risk of gaming is lurking wherever exception reporting is allowed, which was not the case in our program. Nevertheless, a qualitative study in the UK has shown that the use of exception reporting is acceptable.[30] Another unintended consequence mentioned was developing a 'tunnel vision', which means that GPs focus most on those aspects of care that are incentivised. In the UK, performance improvements in those conditions that were not included in the QOF were significantly less than for the incentivised indicators, and these differences increased over time.[31] Focusing on the aspects of care that are incentivised might occur at the expense of other aspects that could be more important, such as personal attention. The results of a study in the UK suggest that the program has changed the nature of the practitioner-patient consultation.[29] Another concern was the data collection, which was very labour intensive. As mentioned before, the stakeholders found a

good compromise in obtaining meaningful information in a feasible manner. The Australian Medical Association also described the high administrative burden that was created by their PIP program.[32] Attaching performance to payments can generate suspicion and resentment, undermining the mutual trust between payer and provider that is essential for any P4P program to operate effectively.[33] About half of the GPs participating in our program did not want to give detailed information about their performance to insurers, for fear that the information might be used for sanctions or penalties.

Methodological considerations

This thesis contains five studies, each of which has specific strengths and limitations, as has been mentioned in the separate chapters. Here we summarize the main methodological considerations.

Overall the development and evaluation of the P4P program can be considered an exploratory study. Attributing the improvements to the P4P program would, for instance, have been easier if the study design had been a randomised controlled trial (RCT). However, most studies on the effect of P4P do not meet RCT criteria.[8] Introducing a control group for such a RCT is difficult because collecting data on performance measurement is an intervention strategy in itself. Also, most practices do not like to be in the control arm of such an experiment. This makes our results responsive to other activities in general practice focused on improving the quality of care of the chosen subjects. However, finding different measures that all changed in the same direction gives us circumstantial evidence that supports the effect of the program.

Practices were able to register voluntarily for this experiment, which may have resulted in overrepresentation of early adopters of the P4P program. It is important that other GPs support the program as well. Whether the designed P4P program reflects the majority of the general practices nationally is uncertain. This might have affected the further implementation of the P4P program. It would be of interest to learn about the experiences of a comparative group of people who were not involved in designing the program and who were not participating as volunteers. It can be argued that this group would probably have been less enthusiastic as a control group in the P4P program.

The link with the national accreditation program for primary care can be considered both a strength and a weakness. The intervention would have been totally bottom-up if we had started from scratch with the domains and indicators for the development of the P4P program. As the stakeholders considered the accreditation program an acceptable foundation for the P4P program, a lot of time was saved in the discussions about design choices, which meant the data collection phase was accelerated.

The time span of the evaluation of the P4P program allowed one post-measure-

ment both for clinical care and patient experience. If we could have done more measurements we would have been able to learn more about the long-term effects of the program. Furthermore, more improvement plans would in all likelihood have been completed after a longer period. Therefore, the conclusions of this study should be interpreted with caution.

Implications

In our project we have obtained indications that our approach to P4P and quality improvement, which is focused on both intrinsic and extrinsic motivation, may be more effective than a system that is focused on either intrinsic or extrinsic motivation. However, we do not know whether this can be sustained over a longer period of time. Therefore, we have presented some of the implications of a national roll out of the P4P program. To a certain extent, the implications are based on some conversations we had with stakeholders several years after finishing the P4P experiment. These stakeholders were representatives of the two financing health insurers and a general practitioner who all actively took part in the former project team. The conversations resulted in hypotheses concerning the national roll out of the P4P program which will be discussed in more detail below. For national roll out, it is important that the program is embedded in daily activities, is not seen as a project that will end and is feasible. Given the development of integrated care, embedding the program in daily activities means that integrated care needs to be incentivised. Furthermore, the program needs to be affordable and arrangements need to be made regionally in which the key figures of the region are involved.

Performance measures for integrated care

In our P4P program, performance measures for general practice were developed and incentivised. In the Netherlands integrated care programs have now been developed for diabetes, COPD and cardiovascular diseases.[34-36] In these programs different healthcare providers are jointly responsible for the care that is delivered, not just the GP and medical specialist but also the physiotherapist, the pharmacist and other allied healthcare providers. Given this development in chronic care, the performance measures in our P4P program should be adapted to a set of performance measures for each healthcare provider and to integrated care as a whole. It is important that the data collection is feasible: having a good infrastructure for data collection is very important, even more so when collecting data for integrated care that is delivered by more than one provider in more than one setting. Therefore, extracting data automatically via the electronic medical record (EMR) system is preferred, like for example that in the UK.[12] The EMR was one of the barriers to national roll out mentioned by the stakeholders. For large scale implementation of the P4P program it is necessary that the EMR is

well organised for retrieving data, and that software to support this process becomes available. Even the early adopters of the P4P program did not necessarily have a good ICT infrastructure; therefore, to get this right, additional investment is needed. Furthermore, the validity and reliability of the data are a concern when extracting data automatically.

When developing performance measures in a participatory way it is a challenge to involve all stakeholders, such as the organisations for dieticians, physiotherapists, pharmacists and so on. Furthermore, it is a challenge to decide who can make the final decisions. We also suggest putting in place a system for revising indicators through which indicators can be deliberately added or removed.

In healthcare in general, and in quality improvement in particular, leadership is very important.[37,38] Research is needed to explore leadership in the field of performance measures for integrated care in a P4P program.

Payment for integrated care

The introduction of a P4P program with performance measures for integrated care has consequences for the payment system. Purchasers wonder how much to pay for quality, which aspects of care should be involved, and what source should be used for this re-investment. The main question is where to find the 'new' money for this kind of experiment. That is why new payment programs have been introduced, especially in the USA. One of the more comprehensive models is the Alternative Quality Contract (AQC) from Blue Cross Blue Shield of Massachusetts (BCBS).[39] The AQC is a contracting model that is based on global payment (capitation) and pay-for-performance. The AQC contains three main features that distinguish it from traditional fee-for-service contracts. Physician groups, in some cases together with a hospital, enter into 5-year global budget contracts (rather than 1-year contracts). AQC groups are eligible for P4P bonuses of up to 10% of their budget, with performance measures for ambulatory care and hospital care each contributing up to half of the calculation of the bonus. AQC groups receive technical support from BCBS, including reports on spending, utilization, and quality, to assist them in managing their budget and improving quality. Such a model suggests management of care on a population level.

A shared saving program could be another alternative. These programs are defined in a fee-for-service system. In these payment models a group of caregivers is made responsible for a certain population (chronically ill, elderly), and they can share payer savings if they meet quality and cost performance benchmarks.[40] Shared savings models can differ in the degree of risk shifting. Some models entail no performance risk to providers even if they experience higher costs or if they do not achieve quality performance goals. Other models require providers to share the financial risk by accepting some accountability for costs

that greatly exceed the goals. The latter gives providers an opportunity to receive proportionately larger bonuses in exchange for this risk.[41] A fundamental question in a shared saving program is the extent to which feasible improvements in performance by themselves can be expected to produce substantial savings. Furthermore high performers with low average costs create value for the system but might not demonstrate 'new' savings.[42]

Regional payment of integrated care

In the Netherlands the payment system for chronic care and prevention benefits from a payment for a certain population within a region when specific criteria for quality are achieved. A combination of the alternative quality contract and shared savings could be a payment system that supports quality improvement in an affordable way. The payments for each region can be used for improving the healthcare in the region with specific agreements about what and how to improve. It is important to involve the key figures in the region to make these agreements. Furthermore, a sound monitoring system is important for all payment systems involving quality parameters and costs. No clear answers can be given about its effectiveness: this payment structure needs to be tested and evaluated scientifically.

Conclusion

We have developed a participatory P4P program for primary care, the results of which were encouraging. New payment models have been suggested, but evidence for them is still lacking. We argue for a systematic evaluation and implementation. Specific lessons can be learned from this experiment, especially since P4P seems to play a prominent role in these new payment models. These lessons concern the positive effect of the involvement of stakeholders in designing the P4P program, the necessity of having a well-organised infrastructure including accessible electronic medical records, and beginning with a national roll out procedure with strong leadership to ensure long-term positive effects.

References

1
Grol R, Wensing MJP, Eccles M. *Improving patient care: the implementation of change in clinical practice*. London, England: Butterworth-Heinemann, 2005.

2
Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, et al. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane database of systematic reviews (Online)* 2012;6:CD000259.

3
Edgman-Levitan S, Dale Shaller PA, McInnes K, Joyce R, Coltin KL, Cleary PD. The CAHPS improvement guide: practical strategies for improving the patient care experience. Harvard: Harvard Medical School, Department of Health Care Policy, 2003.

4
Fairbrother G, Hanson KL, Friedman S, Butts GC. The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates. *Am J Public Health* 1999;89(2):171-75.

5
Hillman AL, Ripley K, Goldfarb N, Weiner J, Nuamah I, Lusk E. The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care. *Pediatrics* 1999;104(4 Pt 1):931-5.

6
Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S. Does pay-for-performance improve the quality of health care? *Ann Intern Med* 2006;145(4):265-72.

7
Rosenthal MB, Frank RG. What is the empirical basis for paying for quality in health care? *Med Care Res Rev* 2006;63(2):135-57.

8
Scott A, Sivey P, Ait OD, Willenberg L, Naccarella L, Furler J, et al. The effect of financial incentives on the quality of health care provided by primary care physicians. *Cochrane Database Syst Rev* 2011;9:CD008451.

9
Van Herck P, De SD, Annemans L, Remmen R, Rosenthal MB, Sermeus W. Systematic review: effects, design choices, and context of pay-for-performance in health care. *BMC Health Serv Res* 2010;10:247.

10
Mannion R, Davies HT. Payment for performance in health care. *BMJ* 2008;336(7639):306-08.

11
Kirschner K, Braspenning J, Jacobs JA, Grol R. Design choices made by target users for a pay-for-performance program in primary care: an action research approach. *BMC Fam Pract* 2012;13(1):25.

12
Roland M. Linking physicians' pay to the quality of care: a major experiment in the United kingdom. *N Engl J Med* 2004;351(14):1448-54.

13
Van den Hombergh P, Schalk-Soekar S, Kramer A, Bottema B, Campbell S, Braspenning J. Are family practice trainers and their host practices any better? Comparing practice trainers and non-trainers and their practices. *BMC family practice* 2013;14:23.

14
Australia M. 2013. Medicare Australia, 2013, http://www.medicareaustralia.gov.au/provider/incentives/pip/index.jsp, year cited 2013, date cited 15 February

15
Association IH. 2013. Integrated Healthcare Association, 2013, http://iha.org/p4p_california.html, year cited 2013, date cited 15 February

16
Damberg CL, Raube K, Williams T, Shortell SM. Paying for performance: implementing a statewide project in California. *Qual Manag Health Care* 2005;14(2):66-79.

17
Employers BN. Quality and Outcomes Framework guidance for GMS contract 2011/12. Delivering investment in general practice. London: UK, 2011.

18
Conrad DA, Christianson JB. Penetrating the "black box": financial incentives for enhancing the quality of physician services. *Med Care Res Rev* 2004;61(3 Suppl):37S-68S.

19
Frolich A, Talavera JA, Broadhead P, Dudley RA. A behavioral model of clinician responses to incentives to improve quality. *Health Policy* 2007;80(1):179-93.

20
Mehrotra A, Sorbero ME, Damberg CL. Using the lessons of behavioral economics to design more effective pay-for-performance programs. *Am.J.Manag.Care* 2010;16(7):497-503.

21
Rosenthal MB, Fernandopulle R, Song HR, Landon B. Paying for quality: providers' incentives for quality improvement. *Health Aff. (Millwood)* 2004;23(2):127-41.

22
Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M. Effects of pay for performance on the quality of primary care in England. *N Engl J Med* 2009;361(4):368-78.

23
Rosenthal MB, Dudley RA. Pay-for-performance: will the latest payment trend improve care? *JAMA* 2007;297(7):740-44.

24
Kirschner K, Braspenning J, Akkermans RP, Annelies Jacobs JE, Grol R. Assessment of a pay-for-performance program in primary care designed by target users. *Family practice* 2012.

25
McDonald R, Roland M. Pay for performance in primary care in England and California: comparison of unintended consequences. *Ann Fam Med* 2009;7(2):121-27.

26
McDonald R, White J, Marmor TR. Paying for performance in primary medical care: learning about and learning from "success" and "failure" in England and California. *J Health Polit Policy Law* 2009;34(5):747-76.

27
Teleki SS, Damberg CL, Pham C, Berry SH. Will financial incentives stimulate quality improvement? Reactions from frontline physicians. *Am J Med Qual* 2006;21(6):367-74.

28
Young GJ, Meterko M, Beckman H, Baker E, White B, Sautter KM, et al. Effects of paying physicians based on their relative performance for quality. *J Gen Intern Med* 2007;22(6):872-76.

29
Campbell SM, McDonald R, Lester H. The experience of pay for performance in English family practice: a qualitative study. *Ann Fam Med* 2008;6(3):228-34.

30
Campbell S, Hannon K, Lester H. Exception reporting in the Quality and Outcomes Framework: views of practice staff – a qualitative study. *Br J Gen Pract* 2011;61(585):183-9.

31
Doran T, Kontopantelis E, Valderas JM, Campbell S, Roland M, Salisbury C, et al. Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework. *BMJ* 2011;342:d3590.

32
Association AM. AMA Submission to the Productivity Commission study on administrative and compliance costs associated with Commonwealth programs that impact specifically on general practice, 2002.

33
Doran T, Roland M. Lessons from major initiatives to improve primary care in the United Kingdom. *Health affairs (Project Hope)* 2010;29(5):1023-9.

34
NDF Zorgstandaard. Transparantie en kwaliteit van diabeteszorg voor mensen met diabetes type 2: Nederlandse Diabetes Federatie, 2007.

35
Zorgstandaard Vasculair Risicomanagement: Platform Vitale Vaten, 2009.

36
Zorgstandaard COPD: Long Alliantie Nederland, 2012.

37
Berwick DM. Continuous improvement as an ideal in health care. *N Engl J Med* 1989;320(1):53-6.

38
Goldberg DG, Mick SS, Kuzel AJ, Feng LB, Love LE. Why do some primary care practices engage in practice improvement efforts whereas others do not? *Health Serv Res* 2013;48(2 Pt 1):398-416.

39
Song Z, Safran DG, Landon BE, He Y, Ellis RP, Mechanic RE, et al. Health care spending and quality in year 1 of the alternative quality contract. *N Engl J Med* 2011;365(10):909-18.

40
Schneider EC, Hussey PS, Schnyer C. Payment Reform. Analysis of Models and Performance Measurement Implications: RAND Corporation, 2011.

41
Institute HCII. Shared saving models, 2013.

42
Weissman JS, Bailit M, D'Andrea G, Rosenthal MB. The design and application of shared savings programs: lessons from early adopters. *Health affairs (Project Hope)* 2012;31(9):1959-68.

Summary

Changing the behaviour of healthcare professionals is necessary to improve quality of care. Audit and feedback is a common intervention that has proven to be effective, although the success rate is sometimes marginal. In a new generation of interventions audit and feedback is extended by setting financial incentives in prospect. These pay-for-performance (P4P) programs have emerged in many forms, but effect studies presented inconclusive results. One of the reasons is probably the difference in designs of the programs compared. A clear framework for the design of a P4P program was lacking in the initial period. Also, the financial incentives themselves varied significantly. A small bonus may have little incentive effect, but a large bonus may provoke negative side effects. Linked to this dilemma is consideration whether P4P programs that are directed more towards raising the intrinsic motivation for quality improvement may enhance their effectiveness. This intrinsic motivation can be stimulated by involving the target users in designing the P4P program. The present thesis focuses on the structural development and evaluation of a participatory pay-for-performance program, in which target users are involved in designing the P4P program.

Chapter 1 describes the objective and the following research questions:

1 What is the impact of feedback on performance accompanied by information on best practices on the accessibility and availability in general practice?
2 Which design choices are made by the target users for a P4P program, in which the different options for performance measurement, appraisal and reimbursement were discussed in a systematic consensus procedure?
3 What is the validity of the clinical care domain?
4 What is the effect of the P4P program on clinical indicators as well as on the patient experience after one year?
5 Has the involvement of target users in designing the P4P program led to positive evaluations and confidence in its future use?

In **chapter 2**, we present a quality-improvement study examining the impact of feedback on performance accompanied by information on best practices on the accessibility and availability in general practice. A pre-intervention and post-intervention data collection was conducted in 36 general practices in the south of the Netherlands. Many patients were not satisfied with the accessibility and availability of general practice, and would like to see improvement. The practices received feedback about their accessibility and availability compared with data from practices of colleagues. The practices developed practice-based improvement plans using these feedback results. Eighty percent of the improvement plans were completed or almost completed in five months. Feedback and practice-based improvement plans were a stimulus to work on and to improve

accessibility and availability. All practices started improvement plans, but the overall effect of the changes was modest.

Chapter 3 describes the choices of the target users for a design of a P4P program in primary care. By adopting a framework for designing P4P programs, a distinction has been made between performance measurement, appraisal and reimbursement. Using an adapted Delphi procedure, the design choices were discussed in two panels of target users. The target users were representatives of 65 general practices and of two health insurance companies in the south of the Netherlands. The panels discussed the performance measures, that is, the domains, subjects and indicators, as well as the period of data collection. The appraisal issues discussed were the unit of assessment (GP, general practice or larger organisational unit), performance standards (absolute versus relative), weighing the domains and indicators, exact calculations, weighing quality level and quality improvement scores, number of levels of quality scores and feedback. Discussion of the reimbursement included the form of remuneration (money, human resources, sabbatical), the size of the bonus and agreements on the spending of the bonus (with or without obligations). It was decided that the performance measures were linked to the Dutch National Accreditation program that was established by The Dutch College of General Practitioners. The performance measures concern three domains: clinical care, practice management and patient experience. The topics presented in clinical care involve mainly indicators that describe chronic care (diabetes care, COPD and asthma care, cardiovascular risk management), some issues on prescription management and uptake rates on influenza vaccination and cervical cancer screening. Practice management and patient experience were measured by validated questionnaires. For the appraisal, the general practice was chosen as unit of assessment. Relative standards were set at the 25th percentile of group performance. The incentive for clinical care was set twice as high as the one for practice management and patient experience. The incentives for clinical care were only related to the process indicators and not to the outcomes. Quality scores were to be calculated separately for all three domains, and for both quality level and quality improvement. The incentive for quality level was set thrice as high as the one for the improvement of performance. For reimbursement, quality scores were divided into seven levels. The bonus aimed at was on average 5% to 10% of the practice income with a maximum of 6890 Euros per 1000 patients. The final design choices were consented to by often a large majority, except for the prescription indicators. The prescription indicators were highly valued by the health insurance representatives, but the GPs questioned these indicators, which resulted in not including the acid suppressive drugs indicators in the program, and including only indicators on prescribing antibiotics. No clear decision could

be made about thresholds of the performance measures. Only a slight majority preferred the relative thresholds. It was argued that relative thresholds would prevent low achievers from dropping out as the benchmark was realistic, and that absolute thresholds could be thought of as outranging their performance capabilities. Overall, this P4P program designed by involving the target users contained elements of the different P4P programs in other countries in a unique combination. The more balanced position of patient experiences in our program is rather exceptional.

In **chapter 4**, we present an observational study based on the medical records of 82 general practices to evaluate the clinical indicators of our P4P program. We evaluated four reliability and validity issues, that is (1) the correlation between the indicator sets, (2) internal consistency of the indicator sets, (3) reliability of indicator scores per patient population and (4) the reliability of the benchmark. The indicators that covered chronic disease management (diabetes, COPD, asthma and cardiovascular risk management), prevention activities (influenza vaccination, cervical cancer screening) and antibiotics policy were correlated weakly, suggesting that the instrument provided a rather broad scope of the practice performance on these issues. Furthermore, the different subjects were each measured by indicators that had sufficient coherence, which suggested that they measured a clear underlying concept. It was calculated that to achieve a reliable indicator score, data from at least 96 patients were necessary when 10% error is allowed. To establish a reliable benchmark 233 practices are needed when 5% error is allowed. The clinical indicators of the P4P program are reliable and valid on the tested properties, but the number of patients and practices needed suggests that public information on single indicators lacks meaning.

Chapter 5 describes the effect of our P4P program on clinical indicators as well as on the patient experience after introducing the participatory P4P program. An observational study was conducted with a pre- and post-measurement in 60 general practices in the south of the Netherlands. The assessment was measured by the indicators on clinical care and patient experience that could have been incentivised according to the practice performance. These measures were process indicators of chronic care (diabetes, COPD, asthma and cardiovascular risk management), indicators of antibiotic prescription and uptake rates (influenza vaccination and cervical cancer screening) as well as indicators of patient experience (GP's functioning and practice organisation). Outcome measures of chronic care were also collected although they were not incentivised. A linear mixed effect model was calculated including the above measures after remuneration based on benchmarking as dependent variables, and the baseline measures, practice type and urbanisation level as covariates. After one year, significant

improvements were observed for the process indicators for diabetes (+10.4%), COPD (+8.1%), asthma (+11.5%) and cardiovascular risk management (+7.9%). Patients' experience of GP's functioning statistically improved by 5.9% and patients' experience of the organisation of care by 5.6%. Five out of seven outcome indicators showed significant improvements, ranging from +5.9% (blood pressure diabetes) to +14.7% (blood pressure CVRM). No significant improvements were seen for influenza vaccination rate and the cervical cancer screening uptake. The clinical process and outcome indicators, as well as patient experience indicators, were affected by baseline performance. Smaller practices showed more improvement. We concluded that a participatory P4P program might stimulate quality improvement in clinical care and improve patient experiences with GP's functioning and the organisation of care.

In **chapter 6**, we present a qualitative study about the experiences of general practices with the P4P program. The involvement of target users in the design choices of a P4P program was supposed to enhance its impact, but little is known about the views of participants in these programs. Thirty out of 65 general practices participating in the P4P program, stratified for bonus, were invited to a semistructured interview on feasibility, feedback of the indicator scores, the bonus, spending of the bonus, unintended consequences and future developments. Content analysis was used to process the resulting transcripts. One practice dropped out as it proved impossible to reach an appointment during the predefined period. Feasibility of the P4P program was questioned due to the substantial time investment. The feedback on clinical care, practice management and patient experience was mostly discussed in the team, and used for improvement plans. One GP considered the feedback as annoying, and for another GP it brought feelings of insecurity. Most practices considered the bonus to be a stimulus to improve quality of care, and as compensation for their effort and time invested. Distinctive performance features were not displayed on, for instance, a website. The bonus was mainly spent on new equipment or teambuilding. As unintended consequences, practices referred to gaming and focusing on those aspects that were incentivised ('tunnel vision'). Speaking about future developments it was suggested changing relative standards into absolute ones. Absolute standards have the advantage that practices can prepare themselves better on expected performance standards. Other proposals included that the audit could profit from more standardisation, and new indicators had to be developed to keep the process going. Linking a part of the bonus to innovation was also suggested. General practitioners considered the feedback and the bonus to be a stimulus for improving quality and performance in the next year, although the data collection involved a substantial time investment. The participants were overall positive about the P4P program, but unintended

consequences like gaming and too narrow a focus on program objectives should be taken into account.

In **chapter 7**, we present the general discussion of this thesis, summarising the main findings of our study, its strengths and weaknesses and main implications for practice and research. We assumed that involving GPs in designing a P4P program in a structured manner would improve intrinsic motivation, and that adding an incentive to performance would stimulate extrinsic motivation. This hypothesis was largely supported by our data. Our P4P program resulted in significant improvements of, on average, 6% in the patient experience and 10% in clinical care. Compared with other programs aimed at either intrinsic or extrinsic motivation, the results of our P4P experiment resulted in greater changes.

In general, the involvement of GPs and representatives of the health insurance companies in designing the P4P program resulted in a program with a smaller number of indicators than in the UK's program (Quality and Outcome Framework), but a larger number of indicators than in programs that run in the US or Australia. The three main differences between our P4P program and other programs are: (1) the use of relative performance standards, (2) rewarding both quality level and performance improvement and (3) the size of the bonus. Despite the quality improvement demonstrated and the fairly positive evaluations of the P4P program by the GPs, national roll out was delayed. In the meantime a national discussion started on integrated primary care. Again, chronic care was chosen as the prior topic in these programs that made use of the same indicators as our P4P program. However, the remuneration was not expressed in a bonus but in bundled payments. In some experimental settings population management will be introduced in which a bonus may play a certain role. The lessons learned from our P4P program showed that the involvement of target users in designing these programs in a structured manner does pay off in the impact on quality of care improvement.

Samenvatting

Om de kwaliteit van zorg te verbeteren, is het veranderen van het gedrag van beroepsbeoefenaren in de gezondheidszorg noodzakelijk. *Audit* en feedback is een veel voorkomende interventie die bewezen effectief is, alhoewel het effect soms marginaal is. In een nieuwe generatie van interventies op het gebied van *audit* en feedback worden consequenties verbonden aan de feedback. Zo wordt bijvoorbeeld een financiële prikkel gekoppeld aan prestaties. Deze *pay-for-performance* (P4P) programma's zijn ontstaan in vele vormen, maar effectstudies laten geen duidelijke resultaten zien. Het aantonen van resultaten wordt bemoeilijkt, omdat de programma's zeer verschillend van inhoud zijn. Een duidelijk kader voor het ontwerp van een P4P programma ontbrak in de eerste periode. Maar ook de omvang van de financiële prikkel zelf verschilde sterk tussen de diverse programma's. Wel bleek dat in het algemeen een lage bonus minder effect had, maar dat een hoge bonus het gevaar in zich draagt van negatieve bijwerkingen. Gekoppeld aan deze discussie kan men zich afvragen of P4P programma's die meer op intrinsieke motivatie zijn gericht de effectiviteit verbeteren. De intrinsieke motivatie kan worden gestimuleerd door het betrekken van de doelgroep bij het ontwerp van het P4P programma. Dit proefschrift richt zich op de gestructureerde ontwikkeling en evaluatie van een *pay-for-performance* programma waarin beoogde gebruikers betrokken zijn bij het ontwerp van het programma.

Hoofdstuk 1 beschrijft het doel en de volgende onderzoeksvragen:

1 Wat is de invloed van feedback én informatie over *best practices* op de toegankelijkheid en de beschikbaarheid in de huisartsenpraktijk?
2 Welke keuzes maken beoogde gebruikers voor de invulling van een P4P programma wat betreft de te meten prestaties, de waardering daarvan en de daarbij passende vergoeding in een systematische consensus procedure?
3 Wat is de validiteit van het domein medisch handelen in het P4P programma?
4 Wat is het effect van het P4P programma op het medisch handelen, alsmede op de patiëntervaringen na een jaar?
5 Heeft de betrokkenheid van de beoogde gebruikers bij het ontwerp van het P4P programma geleid tot positieve evaluaties en vertrouwen in het toekomstig gebruik?

In **hoofdstuk 2** presenteren we het effect van feedback én informatie over de *best practices* op de toegankelijkheid en de beschikbaarheid in de huisartsenpraktijk. Een voor- en nameting werd uitgevoerd in 36 huisartsenpraktijken in het zuiden van Nederland. Veel patiënten waren niet tevreden over de toegankelijkheid en beschikbaarheid van de huisartsenpraktijk en ze zouden graag verbetering zien. De praktijken kregen feedback over hun toegankelijkheid en beschikbaarheid waarbij een vergelijking is gemaakt met praktijken van

collega's. Gebaseerd op de feedback maakten de praktijken verbeterplannen voor hun praktijk waarbij gebruik gemaakt kon worden van *best practices* uit andere praktijken. Tachtig procent van de verbeterplannen werd afgerond of bijna afgerond in 5 maanden tijd. Feedback en praktijkgerichte verbeterplannen waren een stimulans om te werken aan de toegankelijkheid en de beschikbaarheid en het verbeteren hiervan. Alle praktijken zijn gestart met verbeterplannen, maar het overall effect op de toegankelijkheid en de beschikbaarheid was gering.

Hoofdstuk 3 beschrijft de keuzes voor het ontwerp van het P4P programma voor de eerstelijns gezondheidszorg die zijn gemaakt door de beoogde gebruikers van het programma. Bij het opstellen van het P4P programma is systematisch gewerkt. Eerst is gesproken over welke prestaties gemeten worden. Vervolgens is nagegaan hoe deze prestaties worden gewaardeerd. Tot slot is gekoppeld aan deze waardering gesproken over de (hoogte van de) vergoeding. De ontwerpkeuzes werden in twee panels van gebruikers voorgelegd met behulp van een aangepaste Delphi-procedure. De beoogde gebruikers waren 65 huisartsenpraktijken en vertegenwoordigers van de twee zorgverzekeraars in het zuiden van Nederland. De panels bediscussieerden de prestatiemeting, dat wil zeggen de domeinen, de onderwerpen en de indicatoren, alsmede wanneer welke gegevens verzameld worden. De onderwerpen die in het kader van de waardering van de prestaties werden bediscussieerd waren de eenheid van de waardering (huisarts, huisartsenpraktijk of grotere organisatorische eenheid), het wegen van de domeinen en indicatoren, de prestatienormen (absoluut versus relatief), de weging tussen de prestatie op het niveau van de geleverde kwaliteit en de kwaliteitsverbetering, en het aantal niveaus waarin de kwaliteitsscore tot uitdrukking wordt gebracht. Wat betreft de vergoeding is gesproken over de vorm van de beloning (geld, personele ondersteuning, sabbatical), de hoogte van de bonus en de besteding van de bonus (met of zonder verplichtingen). In de consensusbespreking is besloten dat de prestatie-indicatoren worden gekoppeld aan een bestaand instrument van de beroepsgroep, dat gebruikt wordt in de NHG-Praktijkaccreditering®. Deze prestatiemeting bestaat uit drie domeinen: het medisch handelen, praktijkmanagement en patiëntervaringen. De onderwerpen voor het medisch handelen betreffen voornamelijk indicatoren met betrekking tot de chronische zorg (diabetes, COPD en astma, cardiovasculair risicomanagement), voorschrijfgedrag en opkomstcijfers voor griepvaccinatie en baarmoederhalskankerscreening. Praktijkmanagement en patiëntervaringen worden gemeten met gevalideerde vragenlijsten.
Voor de waardering van de prestaties is besloten om net als in de NHG-Praktijkaccreditering® de huisartsenpraktijk als eenheid van de analyse te nemen. Verder is vastgesteld dat er relatieve normen worden gehanteerd waarbij beloning kan plaatsvinden vanaf het 25ste percentiel van de groepsprestatie.

De bonus voor het medisch handelen zal twee keer zo hoog zijn als die voor praktijkmanagement en de patiëntervaringen. De bonus voor het medisch handelen wordt alleen gebaseerd op de procesindicatoren en niet op de uitkomst-indicatoren. Kwaliteitsscores worden voor de drie domeinen afzonderlijk berekend en zowel voor het niveau van de kwaliteit als voor kwaliteitsverbetering. De bonus voor het kwaliteitsniveau zal driemaal zo hoog zijn als die voor de kwaliteitsverbetering. Voor de betaling worden kwaliteitsscores verdeeld in zeven niveaus.
De bonus gaat gemiddeld 5% tot 10% van het praktijkinkomen bedragen met een maximum van 6.890 euro per 1000 patiënten. Deze uiteindelijke keuzes zijn met grote consensus bereikt, met uitzondering van de prestatie over het medicatie-beleid. De prescriptie-indicatoren zijn positief gewaardeerd door de afgevaardigden van de zorgverzekeraar, maar de huisartsen stelden hun vraagtekens bij de inhoudelijke betekenis van deze indicatoren. Dit resulteerde in het niet includeren van de indicatoren over het gebruik van maagzuurremmers in het programma; de indicatoren met betrekking tot het voorschrijven van antibiotica zijn wel in het programma opgenomen. Het meest is er getwijfeld aan het bepalen van een absolute of relatieve prestatienorm. Slechts een kleine meerderheid gaf de voorkeur aan relatieve normen. Doorslaggevend argument voor het hanteren van relatieve normen was het binnenboord kunnen houden van laag scorende praktijken. Gevreesd werd dat absolute normen als niet haalbaar gezien zouden worden door de laag scorende praktijken. Dit P4P programma dat ontwikkeld is samen met de beoogde gebruikers bevat elementen van verschillende P4P programma's uit andere landen die op een unieke manier gecombineerd zijn. De meer evenwichtige positie van patiëntervaringen in ons programma is uitzonderlijk vergeleken met andere programma's.

In **hoofdstuk 4** presenteren we een observationele studie van de medische dossiers van 82 huisartsenpraktijken om de indicatoren van het medisch handelen in ons P4P programma te evalueren. We evalueerden vier betrouwbaarheids- en validiteitsonderwerpen, (1) de correlatie tussen de indicatorensets, (2) interne consistentie van de indicatorensets, (3) betrouwbaarheid van de indicatorscores per patiëntpopulatie en (4) de betrouwbaarheid van de benchmark. De indicatoren met betrekking tot chronische zorg (diabetes, COPD, astma en cardiovasculair risicomanagement), preventieactiviteiten (griepvaccinatie, baarmoederhalskankerscreening) en antibioticabeleid hadden een zwakke correlatie, wat suggereert dat het instrument voorziet in een tamelijk grote reikwijdte qua onderwerpen. Bovendien werden de verschillende onderwerpen elk gemeten met indicatoren die voldoende samenhangen, wat suggereert dat ze een duidelijk onderliggend begrip meten. Om een betrouwbare indicatorscore te behalen is berekend dat gegevens van ten minste 96 patiënten nodig zijn bij

een toegestane fout van 10%. Voor een betrouwbare benchmark zijn 233 praktijken nodig bij een toegestane fout van 5%. De indicatoren van het medisch handelen in het P4P programma zijn betrouwbaar en valide betreffende de getoetste eigenschappen, maar het benodigde aantal patiënten en praktijken geeft aan dat openbare informatie over enkelvoudige indicatoren geen betekenis heeft.

Hoofdstuk 5 beschrijft het effect van ons P4P programma dat in samenspraak met de gebruikers tot stand is gekomen op het medisch handelen en patiëntervaringen na de invoering. Een observationele studie werd uitgevoerd met een voor- en nameting in zestig huisartsenpraktijken in het zuiden van Nederland. Het effect is gemeten voor het medisch handelen en patiëntervaringen waaraan een bonus werd gekoppeld afhankelijk van de praktijkprestaties. De metingen omvatten procesindicatoren voor chronische zorg (diabetes, COPD, astma en cardiovasculair risicomanagement), indicatoren met betrekking tot voorschrijfbeleid van antibiotica en opkomstcijfers (griepvaccinatie en baarmoederhalskankerscreening), alsmede indicatoren met betrekking tot patiëntervaringen (functioneren huisarts en praktijkorganisatie). Uitkomstindicatoren voor chronische zorg zijn ook verzameld, hoewel hier geen bonus aan gekoppeld werd. Met een lineair mixed effect model is het effect berekend van het P4P programma op de indicatoren gedefinieerd in de domeinen medisch handelen en patiëntervaringen, waarbij de metingen voorafgaand aan het uitkeren van de bonus, het praktijktype en de stedelijkheidsgraad als covariaten in het model zijn opgenomen. Na 1 jaar zijn significante verbeteringen gevonden voor de procesindicatoren van de diabeteszorg (+10,4%), COPD-zorg (+8,1%), astmazorg (+11,5%) en het cardiovasculair risicomanagement (+7,9%). Patiëntervaringen met het functioneren van de huisarts verbeterden statistisch met 5,9% en de patiëntervaring met de organisatie van zorg met 5,6%. Vijf van de zeven uitkomstindicatoren lieten significante verbeteringen zien, variërend van 5,9% (bloeddruk diabetes) tot 14,7% (bloeddruk cardiovasculair risicomanagement). Geen significante verbeteringen werden gemeten voor griepvaccinatie en de opkomstcijfers voor baarmoederhalskankerscreening. Kleinere praktijken toonden meer verbetering. We concludeerden dat ons P4P programma kwaliteitsverbetering kan bewerkstelligen in het medisch handelen en dat ook de patiëntervaringen positiever zijn geworden.

In **hoofdstuk 6** presenteren we een kwalitatieve studie naar de ervaringen van huisartsenpraktijken met het P4P programma. Verondersteld werd dat de betrokkenheid van de beoogde gebruikers bij het maken van de keuzes voor het ontwerp van een P4P programma de impact ervan zou vergroten. Onbekend was nog wat de mening van de deelnemers was over ons P4P programma. Dertig van

de 65 huisartsenpraktijken die deelnamen aan het P4P programma, gestratificeerd naar de hoogte van de bonus, zijn uitgenodigd voor een semigestructureerd interview over de haalbaarheid, de feedback, de bonus, de besteding van de bonus, de negatieve bijwerkingen en toekomstige ontwikkelingen. De techniek van inhoudsanalyse is gebruikt voor de uitwerking van de interviews. De geïnterviewden trokken de haalbaarheid van het P4P programma in twijfel vanwege de aanzienlijke tijdsinvestering voor het verzamelen van de benodigde gegevens. De feedback op het medisch handelen, praktijkmanagement en patiëntervaringen is voornamelijk besproken in het team en gebruikt voor verbeterplannen. Eén huisarts vond de feedback vervelend, en voor een andere huisarts bracht het gevoelens van onzekerheid met zich mee. De meeste praktijken zagen de bonus als een stimulans om de kwaliteit van zorg te verbeteren, naast een vergoeding voor hun geïnvesteerde moeite en tijd. Onderscheidende prestaties zijn niet gepubliceerd op bijvoorbeeld een website. De bonus is vooral besteed aan nieuwe apparatuur of teambuilding. Als negatieve bijwerkingen werden *gaming* en focus op de aspecten waar een bonus aan gekoppeld is (‘*tunnel vision*’) genoemd. Als suggestie voor bijstelling van het programma is voorgesteld de relatieve normen te veranderen in absolute normen. Absolute normen hebben het voordeel dat de praktijken zich beter kunnen voorbereiden op de te verwachten prestatienormen. Andere voorstellen gingen over het meer standaardiseren van de *audit* en het ontwikkelen van nieuwe indicatoren om het proces gaande te houden. Het koppelen van een deel van de bonus aan innovatie is tevens gesuggereerd. Huisartsen beschouwden de feedback en de bonus als een stimulans voor het verbeteren van de kwaliteit en prestaties in het komende jaar, alhoewel de dataverzameling een grote tijdsinvestering betrof. De deelnemers waren over het algemeen positief over het P4P programma, maar er dient rekening gehouden te worden met negatieve bijwerkingen zoals *gaming* en een focus op de onderwerpen in het programma.

In **hoofdstuk 7** presenteren we de discussie van dit proefschrift, waarin we de belangrijkste bevindingen van ons onderzoek, de sterke en zwakke punten, en de belangrijkste aanbevelingen voor praktijk en onderzoek samenvatten. We zijn er van uitgegaan dat het betrekken van huisartsen bij het ontwerp van een P4P programma op een gestructureerde manier de intrinsieke motivatie zou verbeteren, en dat het koppelen van een bonus aan prestaties de extrinsieke motivatie zou stimuleren. Deze hypothese wordt grotendeels ondersteund door onze data. Ons P4P programma resulteerde in significante verbeteringen van gemiddeld 6% voor patiëntervaringen en 10% voor het medisch handelen. Vergeleken met andere programma’s die ofwel gericht zijn op intrinsieke dan wel extrinsieke motivatie, laat ons P4P experiment grotere veranderingen zien. Alles bij elkaar genomen heeft het betrekken van huisartsen en vertegenwoordigers van de

zorgverzekeraars bij het ontwerp van het P4P programma geleid tot een programma met minder indicatoren dan in het programma in de UK (Quality and Outcome Framework), maar tot meer indicatoren dan in programma's in de VS of Australië. De drie belangrijkste verschillen tussen ons P4P programma en andere programma's zijn: (1) het gebruik van relatieve prestatienormen, (2) het belonen van zowel het kwaliteitsniveau als de kwaliteitsverbetering, en (3) de hoogte van de bonus. Ondanks de kwaliteitsverbetering die werd aangetoond en de vrij positieve evaluatie van het P4P programma door de huisartsen, is de nationale uitrol vertraagd. In de tussentijd ontstond een nationale discussie over geïntegreerde eerstelijnszorg. In deze programma's ging de aandacht uit naar de chronische zorg en is veelal gekozen voor dezelfde indicatoren als in ons P4P programma. Echter, de vergoeding werd niet uitgedrukt in een bonus, maar in ketenfinanciering. De indicatoren zullen vermoedelijk ook een plek krijgen in een aantal regionale projecten waarbij bekostiging op populatiemanagement wordt geïntroduceerd. De lessen die we geleerd hebben uit de introductie van ons P4P programma kunnen daarbij gebruikt worden. Het betrekken van de beoogde gebruikers bij het ontwerp van deze programma's kan mogelijk ook daar de impact op kwaliteit van zorg verbeteren.

Dankwoord

Jullie mogen best aan me vragen hoe ver ik ben met mijn proefschrift, want het is klaar!
Ik ga proberen iedereen te bedanken die mij, op wat voor manier dan ook, gesteund heeft de afgelopen jaren.

Allereerst wil ik mijn promotor en copromotores bedanken. Richard Grol, bedankt voor je kritische blik en feedback waarmee het onderzoek en de artikelen iedere keer naar een hoger niveau getild werden. Ik heb veel geleerd van al jouw kennis over (het verbeteren van) kwaliteit van zorg en de implementatie van verbeteringen in de praktijk. Jozé Braspenning, wij hebben samen veel zitten puzzelen aan de artikelen. Sommige artikelen kenden aardig wat versies. Samen hebben we de paneldiscussies begeleid, dat vond ik iedere keer weer spannend. Bedankt dat ik altijd bij je binnen kon lopen en voor je steun en vertrouwen de afgelopen jaren. Annelies Jacobs, ik wil je bedanken voor je frisse en heldere blik tijdens het schrijven en jouw kunst van structureren en plannen die ervoor zorgde dat ik het overzicht kon bewaren. Tevens wil ik de (co)-auteurs van de artikelen bedanken, Jako Burgers, Reinier Akkermans en Arna van Doorn.

Toen ik startte als junior onderzoeker ben ik in de wereld van *pay-for-performance* en het project meegenomen door Margot Tacken. Bedankt Margot voor al je hulp en steun zowel live als telefonisch en voor de leuke trip naar Dublin! Voor de (logistiek rondom) dataverzameling en het rekenen met de indicatoren heb ik van meerdere mensen hulp gekregen. Hiervoor wil ik Jan Hermsen, Maarten Krol, Jan van Doremalen, Jolanda van Haren, Yvon de Bruin en Helen Vos bedanken. Angèle van de Ven en Petra Wopereis zijn tot grote steun geweest bij het interviewen van de huisartsen. Jullie ervaring in de huisartsenpraktijk was hierbij onmisbaar. Voor het uitwerken van de interviews wil ik Joost Wammes, Lieke van Brakel, Roosmarijn Jansen en Sophie Verdonschot bedanken.

Gedurende het onderzoek stond er een stuurgroep klaar om mee te denken over de te zetten stappen in de praktijk. Hiervoor wil ik de huisartsen Maarten Klomp en Wim Verstappen bedanken. Tevens wil ik Paul Muijrers, Dennis van de Rijt, Angelique Bonte, Coline van Everdingen en Truus Gootzen van zorgverzekeraar CZ en Jacqueline Batenburg van zorgverzekeraar VGZ bedanken.

Ik bedank ook alle deelnemende huisartsen(praktijken) aan het project die de input hebben geleverd voor de ontwikkeling van het *pay-for-performance* programma en de toetsing hiervan.

Bij IQ healthcare heb ik veel geleerd maar ook veel afleiding gehad in de vorm van gezamenlijk lunchen, film kijken, high tea-en, bowlen, BBQ-en en zelfs skiën in Frankrijk met Selma, Linda, Arna, Lianne, Renate en Lucy. Iedereen bedankt voor deze leuke uitstapjes en gezelligheid. Ik heb een aantal jaar een bijdrage mogen leveren aan de PhD Council. Ik wil de leden bedanken voor de prettige en leuke samenwerking. Lieve Nicole, bedankt voor je enthousiasme, je interesse in mijn wel en wee en je dansje toen je hoorde dat mijn manuscript was ingeleverd. Lieve Anke, bedankt dat je me altijd met open armen hebt ontvangen en voor je kritische Engelse blik!

Dierbare collega's van ZorgImpuls, jullie inspireren me stuk voor stuk. Ik ben blij om in zo een geweldig team te mogen werken. We werken hard, lachen veel en af en toe is er ruimte voor een traan. Lieve (former) BB's, Matine, Annelies en Astrid, jullie hebben me gesteund door dik en dun!

Bedankt voor de TT. Moet je kijken waar die me gebracht heeft! Thanks ladies! Femke, bedankt voor het vertrouwen en de ruimte die je me hebt gegeven om het proefschrift af te ronden.

Ik ben erg trots op mijn twee paranimfen aan mijn zijde. Lieve Elvira, mijn vriendinnetje van de middelbare school, mijn huisgenoot in Maastricht, mijn collega bij IQ, mijn dierbare vriendin waarmee ik keihard kan lachen en serieus kan filosoferen. Onze levens blijven elkaar kruisen. Ik ben dan ook erg benieuwd naar ons volgende kruispunt. Wat een eer dat je mijn paranimf wilt zijn. Lieve Renate, we begonnen in 2006 tegelijkertijd aan onze promotie. Deze gezamenlijke start heeft een bijzondere band geschept. We hebben vaak samen gespard over onze onderzoeksperikelen. Ik heb naast jou mogen staan tijdens jouw promotie en ik ben vereerd dat jij mij nu bij staat.

Lieve Inge, bij jou en Randy is het altijd thuis komen. Bedankt voor je gastvrijheid, je luisterend oor en de warmte die je me geeft. Je scheetjes Mas en Raf geven onze vriendschap een extra dimensie. Lieve Marloes, samen hebben we bijna alle steden in België ontdekt en komen we op de meest leuke en gezellige plekjes uit. Alsof het voorbestemd is. Bedankt voor het kneuteren en keuvelen (dit keer leesbaar). Lieve Lucy, bedankt voor de gezelligheid als kamergenootje bij IQ healthcare en nu als vriendin. Of het nu gaat om een weekendje Leeuwarden, een lunch in Rotterdam, een diner in Arnhem of een statistieksessie, je inspireert me enorm en overlaadt me met positieve energie. Lieve Stefan en Chantal, Alex en ik zijn erg blij met jullie in onze buurt. Samen naar de kermis of tot diep in de nacht uitbuiken van een lekkere BBQ, het kan allemaal. Bedankt daarvoor!

Guus, Mieke, Mariska, Gerwin, Daan en Mark, lieve schoonfamilie, bedankt voor jullie interesse in alle jaren dat ik gewerkt heb aan mijn promotie.

En dan mijn dierbare familie. Als we samen zijn komt onze humor pas écht tot zijn recht, kun je nagaan! Wat kunnen we met elkaar lachen en wat zorgt dat voor een heerlijke ontspanning. Lieve papa en mama, bedankt voor jullie onvoorwaardelijke steun en liefde. Jullie geven me het ultieme thuisgevoel en hebben ervoor gezorgd dat ik weken rustig heb kunnen schrijven. Ook het gezamenlijk vormgeven van het boekje met papa was een geweldige ervaring. Zie hier het resultaat! Lieve Jochen, grote broer, bedankt voor de stevige, warme knuffels en de heerlijke diners en lunches die je bereidt! Lieve Birgit, kleine zus, je bent er altijd voor mij en Alex. Bedankt daarvoor en laten we nog veel lachen, kletsen en gezellige avondjes hebben in 't Appeltje en tegenwoordig ook in de Gouden Griffel. Lieve Birgit en Martine, jullie maken de familie Kirschner compleet en zijn me dierbaar. Bedankt voor jullie interesse en gezelligheid.

Lieve Alex, lieve Xel, wij zijn geen wereldreizigers, maar als wij de kilometers bij elkaar optellen die we de laatste jaren gemaakt hebben om bij elkaar te zijn, zijn we de hele wereld rond geweest! Jij zorgde er iedere keer weer voor dat ik mijn rust vond om aan mijn proefschrift te werken. Bedankt voor je steun en de vrijheid die je me hebt gegeven en voor de schop onder de kont en reflectie die ik (af en toe) nodig had!

Lieve opa, ik draag dit proefschrift aan u op. Wat zou u trots geweest zijn op uw kleindochter.

Curriculum Vitae

Kirsten Kirschner werd geboren op 2 september 1981 in Nijmegen. Ze groeide op in Druten samen met haar broer Jochen en zus Birgit. Ze gaat naar de lagere school in Puiflijk en daarna naar het Pax Christi College in Druten. Ze behaalt haar VWO diploma in 2000.

Kirsten gaat op kamers in Maastricht om aan de Universiteit Maastricht Gezondheidswetenschappen te studeren. Ze kiest voor de afstudeerrichting Zorgwetenschappen. Hier volgt ze verschillende keuzevakken zoals informatie en professionele kwaliteit, transmurale zorg en verslag-legging, gezondheidszorg en politiek, en palliatieve zorg. Tijdens haar afstudeerstage heeft Kirsten bij het Integraal Kankercentrum Limburg (IKL) onderzoek gedaan naar de kwaliteit van zorg in hospicevoorzieningen. Voor dit onderzoek heeft Kirsten enkele hospices in het zuiden van Nederland bezocht en werknemers geïnterviewd.

Na haar studie te hebben afgerond heeft Kirsten enkele maanden bij de thuiszorg gewerkt totdat ze in 2006 bij IQ healthcare aan de Radboud Universiteit Nijmegen begon met haar promotieonderzoek 'Improving primary care by pay-for-performance'.

In 2010 is Kirsten gaan werken bij ZorgImpuls, de regionale ondersteuningsstructuur (ROS) in Rotterdam. Ze verhuisde naar Rotterdam en ging samenwonen met Alexander Odijk. In 2013 zijn Kirsten en Alexander verhuisd naar Berkel en Rodenrijs. Bij ZorgImpuls is Kirsten werkzaam als regioadviseur in de gemeentes Lansingerland, Capelle aan den IJssel en Zuidplas. Tevens is ze aandachtsfunctionaris voor de oefentherapeuten en specialist op het gebied van de ROS-Wijkscan.