Towards optimal test ordering in primary care

Towards Optimal Test Ordering in Primary Care

Wim HJM Verstappen

Promotores:

Prof. dr. R.P.T.M. Grol

Prof. dr. J. M. Grimshaw (University of Ottawa, Canada)

Co-promotor:

Dr. T. van der Weijden

Beoordelingscommissie:

Prof. dr. P.W. de Leeuw, voorzitter

Prof. dr. F. Buntinx

Prof. dr. J.M.A. van Engelshoven

Prof. dr. B.W. Koes (Erasmus Universiteit Rotterdam)

Prof. dr. Th.B. Voorn (Universiteit Utrecht)

Towards Optimal Test Ordering in Primary Care

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
Prof. mr.G.P.M.F. Mols, volgens het besluit van het College van Decanen
in het openbaar te verdedigen op vrijdag 17 september 2004 om 14.00 uur

door *Wim Verstappen*

The study presented in this thesis was conducted at the WOK, Centre of Quality of Care Research Institute Maastricht. The institute participates in CAPHRI (Care and Public Health Research Institute), Maastricht University. At national level it participates in the Netherlands School of Primary Care Research (CaRe), acknowledged in 1995 by the Royal Dutch Academy of Science (KNAW).

Coverdesign: Jaime van Eijkelenborg
Design: Kasper en Pieter van Delft, van delft design
Production: Bookman International b.v.
Coordination: Convention Company

Contents

CHAPTER I Introduction

Wim HJM Verstappen

Introduction

To bridge the gap between evidence-based medicine and practice, we need to learn more about factors and interventions which are important for the implementation of research findings in clinical practice.[1-5] This thesis focuses on factors and interventions that may play a role in the improvement of test ordering behaviour in primary care. The aim of the study reported in this thesis was the systematic development and assessment of an innovative and multifaceted strategy to improve general practitioners'(GPs') test ordering behaviour.

Test ordering in general practice

Test ordering is an important aspect of medical care in general practice. Most GPs in the Netherlands order laboratory, imaging and function tests at the laboratories or imaging and function departments of the regional hospitals. [6-12] GPs themselves can also perform certain laboratory tests, like Hb, ESR, glucose, cholesterol and urinary tests by using desktop analysers available in their own practice.[13] Some GPs also perform function tests, such as ECGs and lung function tests, in their own practice setting.
During the last five years, about a quarter of the Dutch hospitals have set up diagnostic centres for GPs. In some large cities in the Netherlands, GPs can also order tests at regional so-called 'GPs' diagnostic centres' that are not affiliated to a hospital. The files of such diagnostic centres allow data on numbers of tests ordered by GPs to be retrieved, and providing feedback on test-ordering behaviour to the collaborating GPs is one of the main activities of these centres. **Table 1** shows the tests that GPs can order in most of the diagnostic centres. In ordering laboratory tests, GPs collaborating with diagnostic centres use a nationally developed problem-oriented laboratory order form, with all tests grouped in categories of relevant clinical problems; the selection of these tests is based upon national evidence-based guidelines.[14] GPs regularly receive such guidelines for optimal test ordering from the Dutch College of General Practitioners and the national College for Health Insurers.[15]

Over the years, the use of tests has increased in many countries, although inter-doctor variation has been shown to be large.[11 16-20] General practitioners order these laboratory, imaging and function tests for various medical as well as non-medical reasons.[21 22] For instance, GPs may not want to miss important diagnoses or they may want to reassure patients. Test ordering is also important for monitoring chronic diseases or for screening purposes. The increase in the numbers of tests ordered can probably be explained by the ageing of the western population, by rapid advances in diagnostic test technology, by the shifting of care from secondary to primary care, by the growing demand from patients actively asking for tests, by GPs' test ordering routines that are difficult to change and by GPs being more defensive, for fear of making medical mistakes. On the other hand, underuse of diagnostic tests has also been reported.[23 24] These findings indicate that some patients receive sub-optimal care in terms of test ordering, and that there is room for new strategies to achieve improvement.

TABLE 1 TESTS THAT GPS CAN ORDER AT A DIAGNOSTIC CENTRE

LABORATORY TESTS		IMAGING TESTS	FUNCTION TESTS
Alanine aminotransferine	Glucose	Chest X-ray	ECG
Aspartate aminotransferase	HbA1c	X-ray of cervical spine	Exercise ECG
Lactate dehydrogenase	Thyroid stimulating hormone	X-ray of hip	Lung function test
Alkalic phosphatase	Free thyroid hormone	X-ray of knee	IVP
Allergic screening test	Potassium	X-ray of lumbar spine	Gastroscopy
Amylase	Prostate specific antigen	X-ray of shoulder	Sigmoidoscopy
Bilirubin	Serum creatinine	X-ray of skull	
Blood urea nitrogen	Sodium	X-ray of sinus	
C-Reactive protein	Serum uric ucid	Double contrast barium enema	
ESR	γ-Glutamyltransferase	Ultrasound of hepatobiliary tract	
Haemoglobin	Cholesterol	Ultrasound of female genital tract	
Haemoglobin indices	Cholesterol indices	Ultrasound of the kidney	
Haematocrit	Immunoglobulin E		
White blood count			

Determinants of test ordering behaviour

To improve the quality of test ordering behaviour, it is important to gain detailed insight into the determinants of GPs' test ordering behaviour, but it must be admitted that much remains unknown about these determinants.[6 9 19 20 25-40] An improved understanding of these determinants can be used to develop better measures and strategies for change. In everyday medical practice, the decision-making process may be biased by professional-related determinants of test ordering behaviour, such as risk-taking attitudes [41 42] or other personality aspects [29]; knowledge about the appropriate use of tests [43 44] and routines.[45 46] Other determinants of test ordering behaviour are

to be found in the interaction between the professional and his or her direct environment, with one of the factors steering the diagnostic decision-making process being peer influence. Patients' wishes are important as well, as patients have personal views about the value of diagnostic testing.[39] Other contextual determinants have also been reported in the literature.[34,47-51] The availability of test ordering facilities in the region, the way the test ordering procedure is organised, differences in quality improvement programmes, the remuneration system and financial incentives or regulatory sanctions all seem to determine test ordering behaviour in a complex interaction.

To establish determinants of the GPs' actual test ordering behaviour and its variations we decided that it was important to study determinants not only at the individual GP level, as had been done earlier, but also at the level of the local and regional context. These determinants could be used to identify facilitators of and barriers to change, which could be used in designing new strategies.

Effectiveness of strategies to improve test ordering

These facilitators of and barriers to change could help us develop more tailored strategies, as our present knowledge is too limited to decide which strategy would be most effective in improving GPs' test ordering behaviour. Literature reviews have shown that the effectiveness of interventions to influence test ordering has been variable, and results have by no means been unambiguous, due to differences in the type, intensity or setting of the intervention, and methodological differences between studies.[1,4,5,52-57] Some consistent findings have been observed, however. Among the professional-oriented interventions, audit and feedback were effective both in reducing general overuse of tests and in improving the appropriateness of test use according to specific guidelines. Reminders by computer decision support systems seemed to be effective in improving the appropriateness of test use, while organisational interventions proved to influence the general overuse of tests. More studies are required on combinations of professional-oriented and organisation-oriented interventions, e.g. those combining organisational changes, such as changes in the order form, and direct economic incentives for specific test ordering actions. Another promising option is that of interventions using the interaction between the professional and the social network, such as interactive quality improvement meetings in small groups, educational interventions by experts and opinion leaders, and interventions to achieve improvement through patients. It seems desirable to experiment with different combinations of interventions, but it remains hard to predict which combination will be successful. Applying and evaluating the various elements of such interventions separately may reveal the added value of combined strategies.

To add to our knowledge in this field and to evaluate whether our strategy was in line with literature findings, we performed a systematic review of interventions focusing on test ordering behaviour at the request of the EPOC (Effective Practice Organisation Committee) of the Cochrane Collaboration.

A new strategy to improve GPs' test ordering behaviour

On the basis of a preliminary literature study, the council of the Dutch College of Health Insurances recommended the development of a strategy involving feedback and small group quality improvement.[58] In the Netherlands, feedback on test ordering has become a common strategy, with generally positive results, and small group quality improvement sessions within local GP groups have been widely used to discuss prescription behaviour.[59] Local GP groups are an existing part of the infrastructure of Dutch GPs collaborating in a specific region, and sharing patient care outside office hours. Meetings and educational sessions in local GP groups provide a structure for small group quality improvement. An estimated 80-90% of the GPs in the Netherlands meet regularly in their local GP group for some form of continuous medical education. Unfortunately, the effects of this strategy have never been thoroughly assessed.

Based on previous experience with feedback and small group quality improvement and on an overview of the current literature on principles of effective change of clinical performance, we devised a multifaceted strategy.[60-62] This strategy involves a systematic, step-by-step approach, starting with raising awareness of the GP's test ordering performance by individualised, comparative feedback. In the next step, the GPs have to gain a clear understanding of the guidelines on test ordering. Finally, they have to draw up concrete plans for change. Interaction with colleagues can play a role in this process. The combination of feedback, dissemination of evidence-based guidelines and small group quality improvement discussions about the feedback report and the guidelines, within the context of a safe local GP group, is best described as a continuous, systematic and critical reflection by collaborating peers on a GP's own performance and that of others.[61] The hypothesis in our study was that insight into and discussion of one's own performance in a safe group of respected colleagues would be a powerful instrument to improve the quality of test ordering.
Box I gives a detailed overview of our improvement strategy.

The effects of this new strategy were studied in a well-designed experiment. In addition, it was important to evaluate the individual elements of the strategy: were they all necessary or would a simpler intervention suffice? The latter aspect was also important from an economic point of view. Further, it was crucial that such a strategy fits in well with GPs' daily routine. Assessing the actual adoption of the intervention by the GPs required a thorough process evaluation, which could also provide insight into barriers to and facilitators of a large-scale implementation of our strategy. Using existing local GP groups to improve patient care was also in line with the increasing collaboration between GPs in local settings. These local groups are also increasingly becoming parties to agreements with hospitals or to negotiations with health insurers. We expected that making use of such existing structures would make large-scale implementation of the new test ordering strategy relatively easy.

Intervention: the improvement strategy **BOX 1**

The intervention included the following elements: personalised graphical feedback, including a comparison of each GP's own test ordering data with those of colleagues; guideline dissemination and continuous quality improvement meetings in small groups, organised and chaired by the medical coordinator. The strategy was patient care oriented rather than test oriented, in that it did not focus on the volume of specific tests, but on specific clinical problems and associated laboratory, imaging and function tests relevant to everyday GP practice. GPs received three different feedback reports per year on three different clinical problems, together with the national, evidence-based guidelines on test ordering for these specific clinical subjects. This was followed by a 90-minute structured meeting about two weeks later, at which one of the clinical problems was discussed. These small group meetings consisted of three major components. The first was mutual personal feedback by peers, who worked in pairs at the start of the meeting. This was assumed to be a method of peer review that would create a sense of safety. The second component was the introduction and discussion of national guidelines, while the third was the development of individual and group plans for change. This schedule was repeated a year later, using the same three clinical problems, to assess whether a GP or GP group had implemented the plans for change and to initiate further improvements. This iterative aspect was an important feature of the strategy.

Study design

We first studied the determinants of test ordering in a cross-sectional survey of test ordering behaviour among our study population. In addition to characteristics of the professionals and their practice, we were especially interested in the determinants at the local GP group and regional levels. Because the study population was located in various districts and belonged to various local GP groups, multilevel analyses could be performed at the local GP group and regional levels. The next step was the systematic review of interventions focusing on test ordering behaviour. Finally, to determine the effectiveness of our strategy we conducted a randomised controlled trial, at the same time evaluating the strategy's practicability in everyday GP practice. In particular, we evaluated a minimal and a complete variant of the strategy, to determine the added value of the small group quality improvement meetings compared with the feedback only.

Of course, in times of limited resources for health care, costs aspects of new strategies to improve the quality of health care delivery are also important to evaluate. Not only the direct cost effects of such a strategy must be assessed, but also various other costs, such as personnel and co-ordination costs, the time necessary for acquiring the data, analysis and distribution of feedback data and transport costs.[63-67] It is vital to focus not only on the purely financial costs and cost savings, but also on the strain such a strategy puts on the professional in terms of time and energy.

Further, a process evaluation of this quality improvement strategy seems a necessary addition to effect studies to identify important

determinants of change, and to gain insight into barriers to and facilitators of a broader implementation. [68] Finally, evaluating implementation strategies requires a rigorous methodology.[69-74] Randomised experiments with a block design are regarded as powerful instruments in quality improvement research, because they can ensure that non-specific effects are equal in the intervention arms.[75,76] We evaluated whether our design could be applied in other implementation research, and whether it provided a solution to the Hawthorne effect, that is the phenomenon whereby the fact that professionals are taking part in a trial and are being observed may induce them to perform better or more in accordance with what is considered desirable.

Objective and research questions

The main objective of the research project was the systematic development and evaluation of the model for influencing GPs' test ordering behaviour by means of feedback, guidelines and small group quality improvement within local GP groups. We tried to answer the following research questions:

1. a. What is the magnitude of inter-doctor variation in GPs' test ordering behaviour?
   b. Which determinants could explain differences in test ordering by GPs?

2. What is the effectiveness of various strategies to improve doctors' test ordering behaviour: results of a systematic review for the Cochrane Collaboration?

3. What is the effect of a multifaceted strategy on GPs' test ordering behaviour?
   a. What is the effect on the quantity and quality of test ordering by GPs?
   b. What is the surplus value of the complete strategy compared to written feedback only?
   c. What are the costs and cost savings, and what organisational and financial conditions and repercussions are associated with large-scale implementation?

4. Is the strategy applicable in everyday GP practice? Is the improvement strategy actually being implemented in accordance with the protocol and if so, to what degree do the GPs accept it?

5. Can our block design be usefully applied in implementation research, and does it allow non-specific effects, such as the Hawthorne effect, to be controlled for?

Structure of this thesis

The answers to the above-mentioned research questions are presented in this thesis. **Chapter II** presents the results of the survey which was conducted before the intervention. All participating GPs were asked to give their opinion on test ordering within the practice and experience with feedback and quality improvement. The survey was linked to the numbers of tests ordered by the various GPs, derived from the baseline measurement. **Chapter III** provides a systematic literature

review for the Cochrane Collaboration to describe the different approaches and to assess the effectiveness of strategies aimed at influencing test use. In total 98 studies with 118 comparisons were included.
To evaluate (cost) effects, in a clustered trial GP teams were randomised to three arms and they received a quality improvement intervention concerning test ordering on either tests for group A clinical problems (A tests) or tests for group B problems (B tests).(**Figure 1, Table 2**) In all arms the volume of ordering of all A and B tests was monitored. Three 2-armed comparisons were possible. In the trial with the block design we compared the complete intervention in both arms on either the A (arm I) or B tests (arm II); the arms acted as blind controls for each other. In the second trial the complete strategy was compared with a partial strategy. In the classical trial the complete intervention on B tests (arm II) was compared with a control arm without any intervention on B tests (arm III). **Chapter IV** focuses on the outcomes of the two arms in the block design. In this chapter we pay attention to the effects of the total strategy: feedback, group education on guidelines and small group quality improvement. In addition, it discusses the effects of the intervention on various clinical problems, as well as the question whether the numbers of some tests described in the guidelines as 'irrational' had decreased. **Chapter V** describes the second effect evaluation, assessing the added value of small group quality improvement to written feedback after one year of intervention. One arm received feedback as well as taking part in small group quality improvement activities, while the other received feedback only. This chapter also deals in more detail with one of the clinical problems. **Chapter VI** discusses the costs and cost reductions. A real cost-effectiveness analysis was not possible because of the lack of clinical patient data. The chapter discusses

FIGURE 1. STUDY DESIGN

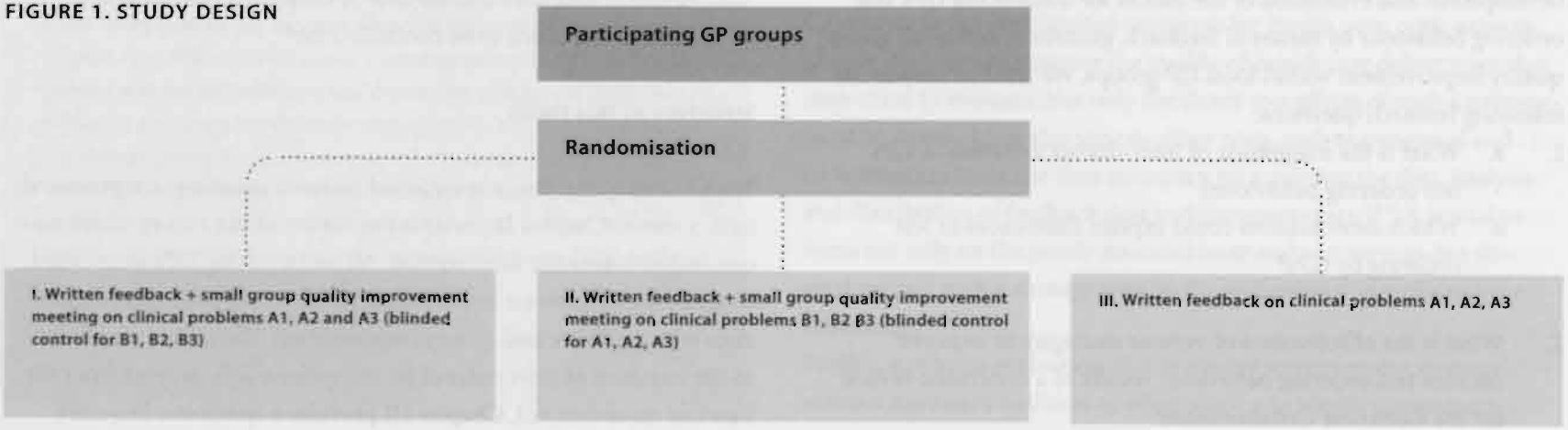


TABLE 2 CLINICAL PROBLEMS AND TESTS USED IN THE TRIAL.

	CLINICAL PROBLEMS / TESTS ARM A		CLINICAL PROBLEMS / TESTS ARM B
A1	**Cardiovascular topics** Cholesterol, subfractions, potassium, sodium, creatinine, (exercise) ECG, BUN	B1	**COPD/Asthma** Allergic screening test, chest X-ray, immunoglobulin E
A2	**Upper abdominal complaints** SGPT, γ-glutamyltransferase, ultrasound scans of hepatobiliary tract, SGOT, LDH, amylase, bilirubin, alkalic phosphatase	B2	**General malaise / Vague complaints** ESR, Hb + indices, Ht, TSH, monospot, leucocyte count
A3	**Lower abdominal complaints** Prostate-specific antigen, CRP, renal ultrasound, IVP, double contrast barium enema, sigmoidoscopy	B3	**Degenerative joint complaints** ESR, uric acid, rheumatoid factors, X-rays of lumbar spine, cervical spine, shoulder, knee, hip

not only costs and cost reductions, but also focuses on a new framework to calculate costs and profits in these types of intervention. The process evaluation is dealt with in **Chapter VII**. Such a process evaluation is regarded as a necessary addition to effect studies to learn about important elements of change, and process data can be very useful for a possible large-scale implementation of the strategy. The block design we used for the effect evaluation is regarded as one of the most powerful designs for investigating quality improvements. **Chapter VIII** pays detailed attention to the method and the study design and evaluates whether or not our study design lived up to expectations. Should this type of design be used more often in future, or do simpler designs suffice? **Chapter IX** presents the general conclusions of the study and the lessons to be learnt from it for the national implementation of this new method. The general conclusion is that the new strategy is an innovative and practicable quality instrument which can be usefully integrated within local and regional quality improvement programmes in an attempt to consistently improve GPs' test ordering behaviour in a practicable, efficient and cost-efficient way.

References

1. Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician performance. A systematic review of the effect of continuing medical education strategies. *JAMA* 1995;274(9):700-5.
2. Bero LA, Grilli R, Grimshaw JM, Harvey E, Oxman AD, Thomson MA. Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. The Cochrane Effective Practice and Organization of Care Review Group. *BMJ* 1998;317(7156):465-8.
3. Shortell SM, Bennett CL, Byck GR. Assessing the impact of continuous quality improvement on clinical practice: what it will take to accelerate progress. *Milbank Q* 1998;76(4):593-624, 510.
4. Solomon DH, Hashimoto H, Daltroy L, Liang MH. Techniques to improve physicians' use of diagnostic tests. A new conceptual framework. *JAMA* 1998;280:2020-2027.
5. Grimshaw JM, Shirran L, Thomas R, Mowatt G, Fraser C, Bero L, et al. Changing provider behavior: an overview of systematic reviews of interventions. *Med Care* 2001;39(8 Suppl 2):Ii2-45.
6. Eisenberg J. Physician utilization. The state of research about physician practice patterns. *Med Care* 1985;213:461-83.
7. Kluijt I, Zaat J, Van der Velden J, Van Eijk J, Schellevis G. Voor een prikje? Het gebruik van klinisch-chemische, hematologische en serologische bepalingen door huisartsen. *Huisarts Wet* 1991;34(2):67-71.
8. Kluijt I, Zaat JOM, Van Eijk JTM, Van der Velden J. Huisarts en beeldvormende diagnostiek. Resultaten uit de Nationale Studie. *Huisarts Wet* 1992;35:188-91.
9. Salloum S, Franssen E. Laboratory investigations in general practice. *Can Fam Physician* 1993;39:1055-61.
10. Van den Bosch W, Bor J, Van de Lisdonk E. Twintig jaar aanvullende diagnostiek in de huisartspraktijk. Cijfers uit de Continue Morbiditeitsregistratie, 1971-1990. *Huisarts Wet* 1993;36(11):365-9.
11. Leurquin P, Van Casteren V, De Maeseneer J. Use of blood tests in general practice: a collaborative study in eight European countries. Eurosentinel Study Group. *Br J Gen Pract* 1995;45(390):21-5.
12. van der Weijden T, van Bokhoven MA, Dinant GJ, van Hasselt CM, Grol RP. Understanding laboratory testing in diagnostic uncertainty: a qualitative study in general practice. *Br J Gen Pract* 2002;52(485):974-80.
13. Hobbs FD, Delaney BC, Fitzmaurice DA, Wilson S, Hyde CJ, Thorpe GH. A review of near patient testing in primary care. [Review] [102 refs]. *Health Technology Assessment* 1997;1(5):1-229.
14. Geldrop W. Een probleemgeorienteerd aanvraagformulier voor laboratorium-onderzoek. Effecten op het aanvraaggedrag van huisartsen. *Huisarts en Wetenschap* 1992;35:192-196.
15. Van Leusden HAIM. *Diagnostisch Kompas 1999/2000*. Amstelveen: Uitgave van het College van zorgverzekeringen (CVZ), 1999.
16. Wertman BG, Sostrin SV, Pavlova Z, Lundberg GD. Why do physicians order laboratory tests? A study of laboratory test request and use patterns. *JAMA* 1980;243(20):2080-2.
17. Kassiser JP. Our stubborn quest for diagnostic certainty. A cause of excessive testing. *New England Journal of Medicine* 1989;320:1489-1491.
18. Woolf SH, Kamerow DB. Testing for uncommon conditions. The heroic search for positive test results. *Archives of Internal Medicine* 1990;150:2451-2457.
19. Ferrier BM, Woodward CA, Cohen M, Goldsmith CH. Laboratory tests: which physicians order more? *Can Fam Physician* 1991;37:347-552.
20. Mabeck CE, Kragstrup J. Is variation a quality in general practice? *Scand J Prim Health Care Suppl* 1993;1:32-5.
21. Winkens R. Improving test ordering in general practice. Rijksuniversiteit Limburg, 1994.
22. Dijksterhuis PH, Van Boven C. De schatbare waarde van aanvullende diagnostiek. *thesis* 1995.
23. Miller WL, McDaniel RR, Jr., Crabtree BF, Stange KC. Practice jazz: understanding variation in family practices using complexity science. *J Fam Pract* 2001;50(10):872-8.
24. Woolf SH, Rothemich SF. Overuse of administrative data to measure underuse of care. *JAMA* 2001;285(6):736-7.
25. Hemenway D, Killen A, Cashman SB, Parks CL, Bicknell WJ.

Physicians' responses to financial incentives: evidence from a for-profit ambulatory care centre. *N Engl J Med* 1990;322:1059-1063.

26. Grol R, Whitfield M, De Maeseneer J, Mokkink H. Attitudes to risk taking in medical decision making among British, Dutch and Belgian general practitioners [see comments]. *Br J Gen Pract* 1990;40(333):134-6.
27. Moskowitz AJ, Kuipers BJ, Kassirer JP. Dealing with uncertainty, risks, and tradeoffs in clinical decisions. A cognitive science approach. *Ann Intern Med* 1988;108(3):435-49.
28. Bugter Maessen AM, Winkens RA, Grol RP, Knottnerus JA, Kester AD, Beusmans GH, et al. Factors predicting differences among general practitioners in test ordering behaviour and in the response to feedback on test requests. *Fam Pract* 1996;13(3):254-8.
29. Ornstein SM, Markert GP, Johnson AH, Rust PF, Afrin LB. The effect of physicians personality on laboratory test ordering for hypertensive patients. *Med Care* 1988;26:536-543.
30. Hjortdahl P, Borchgrevink CF. Continuity of care: influence of general practitioners' knowledge about their patient on use of resources in consultations. *BMJ* 1991;303:1181-1184.
31. Holtgrave DR, Lawler F, Spann SJ. Physicians'risk attitude, laboratory usage, and referral decisions: the case of an academic family practice center. *Med Decis Making* 1991;11:125-130.
32. Kikano GE, Stange KC, Flocke SA, Zyzanski SJ. Effect of the white blood count on the clinical management of the febrile infant [see comments]. *J Fam Pract* 1991;33(5):465-9.
33. Durand Zaleski I, Rymer JC, Roudot Thoraval F, Revuz J, Rosa J. Reducing unnecessary laboratory use with new test request form: example of tumour markers. *Lancet* 1993;342(8864):150-3.
34. Rink E, Hilton S, Szczepura A, Fletcher J, Sibbald B, Davies C, et al. Impact of introducing near patient testing for standard investigations in general practice see comments]. *BMJ* 1993;307(6907):775-8.
35. Royal College of Radiologists Working Party. Influence of Royal College of Radiologists' guidelines on referral from general practice. *BMJ* 1993;306:110-1.
36. Hagen MD. Test Characteristics. How good is that test? *Prim Care* 1995;22(2):213-33.
37. Koide D, Ohe K, Kitamura K, Kitagawa M, Yoshihara H, Nagase T, et al. System for warning on excessive laboratory tests. *Japan Journal of Medical Informatics* 1995;15(4):217-227.
38. Cranney M, Walley T. Same information, different decisions: the influence of evidence on the management of hypertension in the elderly. *Br J Gen Pract* 1996;46(412):661-3.
39. McDonald IG, Daly J, Jelinek VM, Panetta F, Gutman JM. Opening Pandora's box: the unpredictability of reassurance by a normal test result [see comments]. *BMJ* 1996;313(7053):329-32.
40. Valenstein P. Managing physician use of laboratory tests. *Clinics Laboratory Medicine* 1996;16:749-772.
41. Holtgrave DR, Lawler F, Spann SJ. Physicians' risk attitudes, laboratory usage, and referral decisions: the case of an academic family practice center. *Med Decis Making* 1991;11(2):125-30.
42. Zaat JOM, Eijk JTM. General practitioners' uncertainty, risk preference, and use of laboratory tests. *Med Care* 1992;30:846-854.
43. Hicks RJ, Hamm RM, Bemben DA. Prostate cancer screening. What family physicians believe is best. *Arch Fam Med* 1995;4(4):317-22.
44. Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G. Medicine. Communicating statistical information. *Science* 2000;290(5500):2261-2.
45. Little P, Cantrell T, Roberts L, Chapman J, Langridge J, Pickering R. Why do GPs perform investigations?: The medical and social agendas in arranging back X-rays. *Fam Pract* 1998;15(3):264-5.
46. Shye D, Freeborn DK, Romeo J, Eraker S. Understanding physicians' imaging test use in low back pain care: the role of focus groups [see comments]. *Int J Qual Health Care* 1998;10(2):83-91.
47. Zaat JOM, Van Eijk JT, Bonte HA. Laboratory test form design influence test ordering by general practitioners in the Netherlands. *Med Care* 1992;30:189-198.
48. Zaat JO, Schellevis FG, van Eijk JT, van der Velden K. Do out-of-office laboratory tests affect diagnoses in general practice? *Scand J Prim Health Care* 1995;13(1):46-51.
49. van Walraven C, Goel V, Chan B. Effect of population-based interventions on laboratory utilization: a time-series analysis. *JAMA* 1998;280(23):2028-33.
50. Delaney BC, Hyde CJ, McManus RJ, Wilson S, Fitzmaurice DA, Jowett S, et al. Systematic review of near patient test evaluations in primary care. *BMJ* 1999;319(7213):824-7.

51. Van der Weijden T, Grol R, Winkens R, Buntinx F, ter Riet G, Klazinga N. Interventions aimed at influencing the use of diagnostic tests. The relevance of attention for contextual factors.[Protocol]. *Cochrane Library* 2001.
52. Mugford M, Banfield P, O'Hanlon M. Effects of feedback of information on clinical practice: a review. *BMJ* 1991;303:398-402.
53. Axt Adam P, van der Wouden JC, van der Does E. Influencing behavior of physicians ordering laboratory tests: a literature study. *Med Care* 1993;31(9):784-94.
54. Oxman AD, Thomson MA, Davis DA, Haynes RB. No magic bullets: A systematic review of 102 trials of interventions to improve professional practice. *Can Med Ass J* 1995;153:1423-1431.
55. Thomson O'Brien MA, Oxman AD, Davis DA, Haynes RB, Freemantle N, Harvey EL. Audit and feedback: effects on professional practice and health care outcomes. *Cochrane Library* 1997.
56. Thomson O'Brien MA, Oxman AD, Davis DA, Haynes RB, Freemantle N, Harvey EL. Audit and feedback versus alternative strategies: effects on professional practice and health care outcomes. *Cochrane Library* 1997.
57. Thomson O'Brien MA, Oxman AD, Haynes RB, Davis DA, Freemantle N, Harvey EL. Local opinion leaders: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev* 2000(2):Cd000125.
58. Grol R, van der Weijden T, Wensing W, te Giffel M. Effecten van methoden in interventies om richtlijnen voor goede zorg in te voeren en het professioneel handelen te beinvloeden. Eindrapport van de voorstudie. Nijmegen/Maastricht: WOK, 1996.
59. Terra R, Vermuë R, de Kroon A, Kolker L. *Samenwerking huisarts-apotheker. Een werkboek voor farmacotherapie-overleg.* Utrecht: Stichting O&O, 1989.
60. Lawrence M, Schofield T. *Medical audit in primary health care.* Oxford: Oxford University Press, 1993.
61. Grol R, Lawrence M. *Quality improvement by peer review.* Oxford: Oxford University Press, 1995.
62. Fraser R, Lakhani M, Baker R. *Evidence-based audit in general practice.* Oxford: Reed Educational and Professional Publishing Ltd, 1998.
63. Gold M. *Cost-effectiveness in health and medicine*, 1997.
64. Mason J, Eccles M, Freemantle N, Drummond M. A framework for incorporating cost-effectiveness in evidence-based clinical practice guidelines. *Health policy* 1999;47(1):37-52.
65. Sculpher M. Evaluating the cost-effectiveness of interventions designed to increase the utilization of evidence-based guidelines. *Fam Pract* 2000;17(31):1s26-31.
66. Mason J, Freemantle N, Nazareth I, Eccles M, Haines A, Drummond M. When is it cost-effective to change the behavior of health professionals? *JAMA* 2001;286(23):2988-92.
67. Brown CA, Belfield CR, Field SJ. Cost effectiveness of continuing professional development in health care: a critical review of the evidence. *BMJ* 2002;324(7338):652-5.
68. Hulscher ME, Laurant MG, Grol RP. Process evaluation on quality improvement interventions. *Qual Saf Health Care* 2003;12(1):40-6.
69. Campbell DT, Stanley J. *Experimental and quasi-experimental design for research.* Chicago: Rand McNally, 1966.
70. Stephenson J, Imrie J. Why do we need randomised controlled trials to assess behavioural interventions? *BMJ* 1998;316(7131):611-3.
71. Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. The implications of adopting a cluster design are still largely being ignored. *BMJ* 1998;317(7167):1171-2.
72. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;321(7262):694-6.
73. Grimshaw J, Campbell M, Eccles M, Steen N. Experimental and quasi-experimental designs for evaluating guideline implementation strategies. *Fam Pract* 2000;17 Suppl 1:S11-6.
74. Grol R, Baker R, Moss F. Quality improvement research: understanding the science of change in health care. *Qual Health Care* 2002;11(2):110-1.
75. Parsons HM. What Happened at Hawthorne? New evidence suggests the Hawthorne effect resulted from operant reinforcement contingencies. *Science* 1974;183:922-932.
76. Adair JG, Sharpe D, Huynh C-L. Hawthorne Control Procedures in Educational Experiments: A Reconsideration of Their Use and Effectiveness. *Review of Educational Research* 1989;59(2):215-228.

CHAPTER II

Variation in test ordering behaviour of general practitioners: professional or context-related factors?

Wim HJM Verstappen
Gerben ter Riet
Willy I Dubois (†)
Ron Winkens
Richard PTM Grol
Trudy van der Weijden

Published in Family Practice 2004;21:385-93

Abstract

Objective

To describe GPs' test ordering behaviour, and to establish professional and context-related determinants of GPs' inclination to order tests.

Design

Cross-sectional analysis of the combined number of 19 laboratory and 8 imaging tests ordered by GPs, collected from five regional diagnostic centres. In a multivariable multilevel regression analysis, these data were linked with survey data on professional characteristics such as knowledge about and attitude towards test ordering, and with data on context-related factors such as practice type or experience with feedback on test ordering data.

Setting

229 GPs in 40 local GP groups from five regions in the Netherlands.

Main outcome measure

Percentage point differences associated with professional and context-related factors.

Results

Total median number of tests per GP per year was 998 (interquartile range 663 to 1500), with significant differences between the regions. The response to the survey was 97 %.

At professional level 'individual involvement in developing guidelines' (yes versus no), and at context-related level 'group practice' (versus single-handed and two-person practices) and 'more than one year of experience working with a problem-oriented laboratory order form' (yes versus no) were associated with 27%, 18%, and 41% lower numbers of tests ordered, respectively.

Conclusion

In addition to professional determinants, context-related factors appeared to be strongly associated with the numbers of tests ordered. Further studies on GPs' test ordering behaviour should include local and regional factors.

Key words

Family practice, utilisation, physician's practice patterns, test ordering, inter-doctor variation, quality assurance; health care.

Introduction

The use of laboratory and imaging tests by general practitioners (GPs) is increasing in many countries and inter-doctor variation has been shown to be large.[1-3] The reasons for the increase in the numbers of tests ordered are still imperfectly understood, and probably complex. Possible explanations include the expansion of modern diagnostic technology, increased fear of litigation and lack of knowledge about appropriate test use.[4-6] Furthermore, monitoring of chronic diseases is increasingly performed by GPs, due to a shift of care from hospital to primary care.[7]

Improving the quality of test ordering requires a thorough understanding of the causal determinants of test ordering behaviour.[8-11] Previous studies into determinants of test ordering have, in general, yielded inconsistent conclusions. Various professional or practice-related factors have been held responsible for the inter-doctor variation (GP's age, years of experience as a GP, GP's attitude towards risk-taking, practice size and practice type), but no single determinant has been found to be very influential across all of these studies.[12-18] The present study attempted to investigate the influence of context-related determinants not only at practice level but also at the level of local GP groups, such as differences between GP groups in patterns of collaboration, and at the regional level, such as differences between regions in quality improvement programmes or ways of organising test requests. We studied the variation in actual test ordering behaviour among a large group of GPs, to assess determinants of inter-doctor variation, at both the professional level and the level of the local and regional context.

Methods

Design and population

We performed a cross-sectional study of the numbers of tests ordered by GPs, and linked these test ordering data with data from a survey among the study population. Test data were retrieved from the files by staff members of five participating diagnostic centres. A diagnostic centre is an institute, usually associated with a hospital, where GPs can order tests without referring the patient to an outpatient clinic. One of the tasks of the medical coordinator of such a centre is to provide feedback to the GPs about their test ordering. The five different diagnostic centres included in the study used similar problem-oriented test ordering forms for laboratory tests with tests categorised into groups based on clinical problems. The study population consisted of GPs associated with these regional diagnostic centres and whose individual test ordering data could be retrieved. Dutch GPs collaborate with colleagues in so-called local GP groups. They share patient care outside office hours and most groups provide continuing medical education as an important activity. GPs consented to having their individual data on test ordering behaviour used for research purposes.

Variables and instruments

a) The dependent variable for the multivariable regression analysis was the total number of tests that the GP requested in one year (1997). Data of 27 tests (19 laboratory and 8 imaging) were retrieved (**Table 1**). Data on the desktop tests that many GPs regularly perform in their own practice (ESR, haemoglobin, glucose and cholesterol) could not be retrieved, and these tests were therefore excluded.

b) The GPs in the study population were surveyed on the following professional and context-related determinants:

-Professional characteristics: age, number of years of experience, working full time (5 days) or part time, knowledge of diagnostic accuracy measures e.g. sensitivity, predictive value, involvement in guideline development and personal opinions on test ordering. The latter variable was measured on a five-point scale, with options ranging from disagree to agree.

- Context-related determinants: At practice level, we determined practice type, size and location of practice, fraction of privately insured patients (compared to sick fund-insured patients)[16], the fraction of patients older than 65, level of computerisation, distance to the laboratory and imaging facility, and use of desktop equipment. Use of desktop equipment was measured on a four-point scale ranging from never to always. At the local GP group level, we measured quality improvement activities in the GP group setting (yes/no), presence of at least one member who participated (or had participated) in guide-

TABLE 1 TESTS RETRIEVED FROM DIAGNOSTIC CENTRES

LABORATORY TESTS			IMAGING TESTS
Packed Cell Volume	Uric acid	Bilirubin	Chest X-ray
White Blood Count	Prostate Specific Antigen	Immunoglobulin E	Double contrast barium enema
C-Reactive Protein	Alanine Aminotransferase	Allergic screening test	Ultrasound of hepatobiliary tract
Thyroid Stimulating Hormone	Aspartate Aminotransferase		X-ray of cervical spine
Potassium	γ-Glutamyltransferase		X-ray of lumbar spine
Creatinine	Alkalic Phosphatase		X-ray of hip
Blood Urea Nitrogen	Lactate Dehydrogenase		X-ray of knee
Sodium	Amylase		X-ray of shoulder

line development for the Dutch College of General Practitioners (yes/no) and presence of a joint strategy on medication and test ordering in the local GP group (yes/no). At the regional level, we assessed the experience with feedback from the regional diagnostic centre (yes/no) and whether respondents had at least one year experience with the problem-oriented laboratory form (yes/no).

Analysis

Descriptive analyses were performed on test ordering data relating to the 27 tests selected, both for all 27 and for laboratory and imaging tests separately; differences in test ordering data between regions were tested with the Kruskal-Wallis test. To obtain a normal distribution of the dependent variable, all regression analyses were performed with the log-transformed total number of tests ordered. As a consequence, $e^{\text{regression coefficient}}$ reflects a relative risk, and results are reported as percentage point changes associated with the various independent variables.

As an initial step in the regressions analysis, we first conducted a stepwise backward linear regression analysis for each region separately. This approach shows which variables predict best the number of test orders for each region. In these analyses, all variables are initially entered into the model. The regression algorithm then removes – taking into account the effects of others – those variables that do not have a strong independent association with the number of test ordered. Using robust variance estimation, we took into account that – even within the same region – the numbers of test orders GPs' requested cannot be assumed to be statistically independent from each other, because the test ordering behaviour of two GPs within the same GP group may be more similar than that of two GPs from different GP groups. In this initial step of the regression analyses we adjusted for working full time or part time, and the practice size, that is, these variables were forced into the model and were never omitted. The effect of any other variables should be seen in the context of these two. In accordance with the statistical literature, the p-values for entry into or removal from the multivariable model were set at 0.15, and 0.20. In an effort to avoid the selection of too many variables and over-fitting of the data set, only those variables that were selected in each region by this stepwise procedure were eligible for entry into the multilevel multivariable analysis.

In the final regression model, the data had a clear hierarchical structure, with GP groups operating under single regional diagnostic centres and GPs collaborating within GP groups. Again, one should assume that test ordering behaviour of two GPs within the same GP group may be more similar than that of two GPs from different GP groups. The same holds for GP groups within a region being perhaps more similar than two GP groups randomly chosen from different regions. Therefore, the data were modelled in a three-level multilevel analysis model using the Stata command *gllamm* (Generalized Linear Latent and Mixed Models) with GP group and region as the random coefficients. Eligible for the multilevel model were the variables selected by the previously described stepwise procedure for each region separately. In addition, all context-related factors measured at local GP group and regional level, were entered.

Thus, the initial multilevel model contained 11 independent determinants (**Table 3**). To adjust for practice size, the natural logarithm of practice size was entered as an offset variable.[19] Briefly, this was done because it was the number of tests ordered that was essential, rather than the order rate, that is, the number of orders per potential patient who triggered the order by his or her visit to the GP. No tests for interactions were performed to avoid the risk of false-positive associations in subgroups before the theoretical mechanisms underlying test ordering are better understood. The likelihood ratio test was used to decide which levels would be retained. All analyses were carried out using Stata statistical software (Release 7.0. College Station, TX: Stata Corporation).

Results

Individual test ordering data were retrieved for 229 GPs, working in 40 local GP groups in the five selected regions in the Netherlands (**Table 2**). **Figure 2** demonstrates the large variation between regions in the total number of tests ordered ($p<0.001$). In region III, the median number of tests ordered proved to be more than twice that in region II. Of the 229 GPs, 221(97 %) returned the questionnaire. Compared with all Dutch GPs, the study population included more male GPs and more GPs working in urban practice locations. Two-person practices were underrepresented while relatively more GPs practised in group practices. (Data not shown) **Table 3** presents some characteristics of the study population at GP, practice, and local GP group levels. Eighteen GPs were actually involved in developing guidelines.

A knowledge question, involving the application of Bayes' theorem to a patient case, was correctly answered by 16% of the study population. One hundred and eleven GPs (55%) answered that they would feel uncomfortable if it appeared that they clearly ordered more tests than their colleagues. By contrast, nine GPs (4.1%) would be uncomfortable if they ordered fewer tests. There was a desire to discuss personal test ordering behaviour in local GP groups, and to receive feedback on test ordering from the diagnostic centre. At the local group and regional levels, 22 local GP groups had experience of discussing their test ordering behaviour in the local GP group, which had led to (group) plans for change. At regional level, there was only one region (region I) where the diagnostic centre was already providing individualised feedback on test ordering behaviour, while two of the five regions had introduced the problem-oriented form more than one year ago (regions I and II).

Determinants of test ordering variation

Table 3 also shows the professional and context-related variables that were eligible for entry in the multilevel model. The variable location of practice, whose omission had a negligible effect on the coefficients of the remaining variables, was omitted. The random variation due to the local GP group level proved to be small and insignificant after the three GP group level variables had been omitted. Therefore, the local GP group level was omitted, and our final multilevel model contained 7 variables. Our final two-level model explained about 30% of the variation in test ordering. Two of the variables of the final multilevel model were at the professional level: working full time or part time,

TABLE 2 DISTRIBUTION OF NUMBERS OF TEST ORDERED BY 229 GPS IN FIVE REGIONS

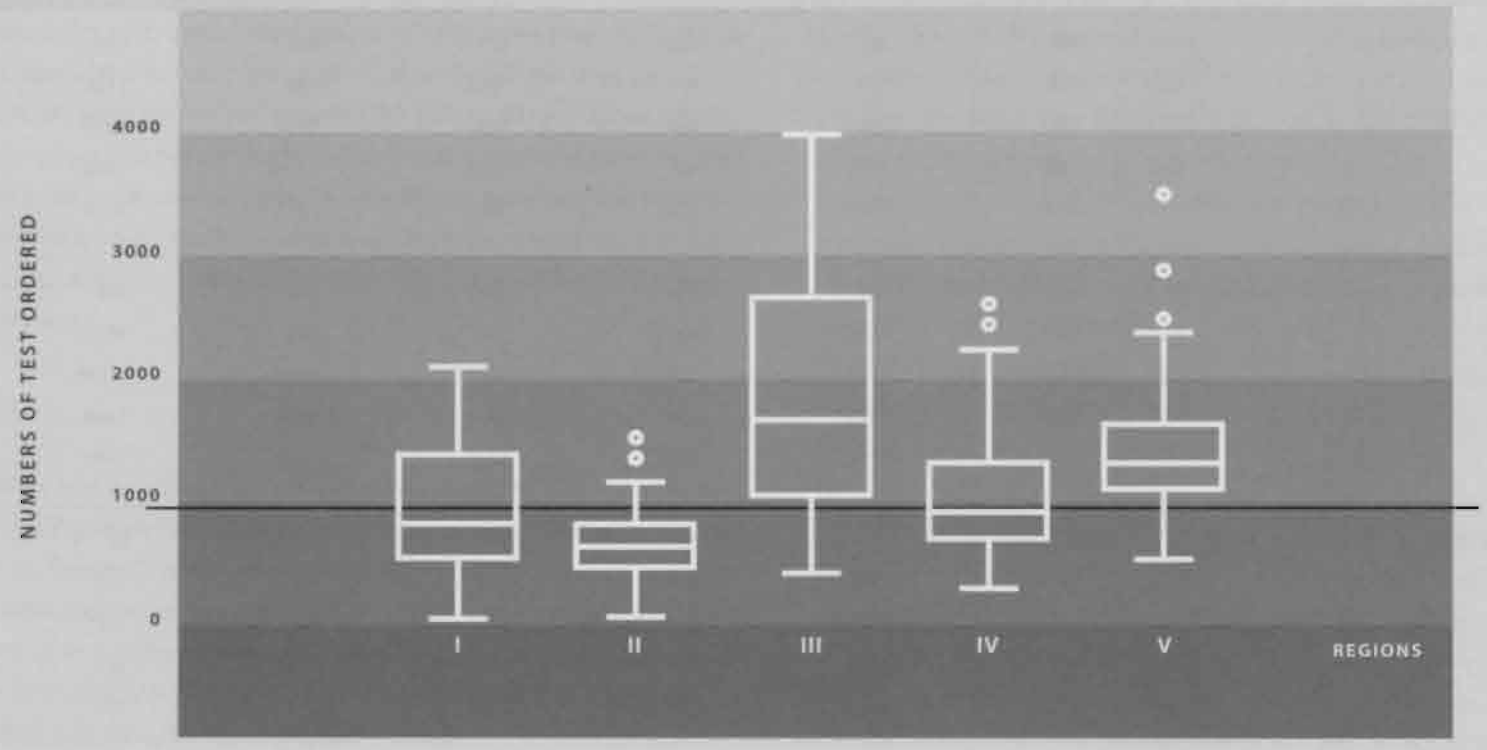
		TOTAL	REGION I	REGION II	REGION III	REGION IV	REGION V
Total numbers of tests ordered	**P5**	364	261	322	617	349	577
	P25	663	576	499	1085	694	1125
	P50	998	860*	666*	1742*	891*	1273*
	P75	1500	1436	847	2781	1344	1608
	P95	2648	1960	1293	3805	2413	2674
Total numbers of laboratory tests ordered	**P5**	303	157	250	498	332	448
	P25	565	456	400	942	569	903
	P50	839	691*	568*	1469*	799*	1078*
	P75	1271	1116	730	2498	1194	1398
	P95	2297	1732	1104	3445	2071	2249
Total numbers of imaging tests ordered	**P5**	40	57	34	74	35	38
	P25	99	110	61	173	96	128
	P50	146	159*	90*	243*	142*	162*
	P75	218	245	132	316	175	221
	P95	370	470	254	379	382	383

*= p<0.001 Kruskal Wallis

P represents *percentile* of the distribution. For example P25, means that 25% of all values are lower than this value.

P50 is identical to the median.

FIGURE 1. BOX PLOT SHOWING THE DISTRIBUTION OF THE NUMBERS OF LABORATORY AND IMAGING TEST ORDERED BY 229 DUTCH FAMILY PHYSICIANS IN EACH OF FIVE REGIONS IN 1997.



The horizontal line shows the overall median number of tests (998) ordered. The horizontal lines within the boxes represent the medians for each respective region. The lower and upper ends of the boxes are the lower and upper quartiles. The 'antennas' sticking out from the boxes delineate where 95% of the observations lie. Dots represent the number of tests ordered by physicians who ordered extremely many tests compared to colleagues within their region. The graph shows the large interregional differences with respect to the average number of test orders as well as with respect to the variation in the numbers of tests ordered. For example, 50% of physicians in region II ordered between 499 and 847 tests, whereas these numbers are 1085 and 2781 for the physicians in region III.

and participation in the production of a guideline. Three variables were at context-related practice level: type of practice, distance to an imaging facility, and distance to a laboratory facility. Two variables were at context-related regional level: feedback on test ordering and experience with the problem-oriented form. **Table 4** shows detailed results of the final two-level model. At the professional GP level, having been actively involved in national guideline setting was associated with a 27% lower volume of tests ordered compared with non-active GPs. The practice type contributed significantly to the variation in test ordering: GPs working in group practices ordered about 18% fewer tests than those in single-person or two-person practices. At context- related regional level, having had at least one year of experience with the problem-oriented laboratory form was associated with a 41% lower volume of tests ordered. The intra-class correlation coefficient at region level was 0.304, meaning that the variation between regions was large compared to the variation within regions, which supports the assumption that variability in test ordering is strongly correlated with a region factor.

Discussion

To our knowledge, the present study is the first to explicitly include context-related variables at GP group and regional level. This enabled us to focus on the variation in GPs' test ordering behaviour in relation to both professional and context-related determinants. We found, to our surprise, a large variation in test ordering between the regions, and we determined three variables that were independently and strongly associated with the volume of tests, namely involvement in developing guidelines, working in a group practice, and having had more than one year of experience with a problem-oriented form.

At the level of the professional, GPs who were involved in developing national clinical guidelines (in the context of the Dutch College of GPs programme for guideline setting) ordered clearly fewer tests than other GPs. Although this subgroup represents a minor and probably selected proportion of the GPs, discussing guidelines and the underlying medical evidence might be an important part of a strategy to improve test ordering behaviour.[20-23] Secondly, at context-related practice level, working in a group practice was associated with a considerably lower number of tests ordered as well. This finding, which probably results from general discussions of and reflections on practice behaviour in such group practices, is in line with earlier findings related to prescription behaviour.[24-26] Finally, at the regional level, it was particularly the level of experience with a problem-oriented test ordering form that appeared to have a large impact on the numbers of tests ordered. It is not so much the influence of the order form itself that is surprising, but rather the magnitude of this effect.[27,28] The present study was unable to explain all of the interregional variation. Of course, disease-related factors are also important in the variation of test ordering. Although there might be slight differences in morbidity between the regions, it is unlikely that differences in case-mix play an important role, because a total of about 550.000 patients were involved. Explaining this interregional variation will require more research, which should include patient-related, organisational and socio-cultural determinants.

TABLE 3 INDIVIDUAL AND CONTEXT-RELATED DETERMINANTS OF THE NUMBER OF TESTS ORDERED BY 221 GPS IN 1997

STUDY POPULATION CHARACTERISTICS	ELIGIBLE FOR INITIAL MULTILEVEL MODEL		B	SE B	TOTAL
DETERMINANTS RELATED TO GPS (N = 221)					221
Male					191
Age (SD)					46.1 (6.2)
GP's number of years of experience in years (SD)					15.5 (7.6)
	Work time factor	5 days	Reference		
		> 4½ days	-0.0756	0.1201	169
		4 days	0.2333	0.1158	25
		< 4 days	-0.1031	0.098	26
	Involved in developing guidelines		-0.2300	0.1269	18
GPs answering questions on diagnostic accuracy correctly					16
Don't want to order more tests than colleagues (scale 1-5)*					3.2
Desire to discuss test ordering in local groups (scale 1-5)*					4.1
Desire to receive feedback on test ordering (scale 1-5)*					4.1
Attitude to risk taking (scale 1-5)*					2.7
Desire to have direct access to MRI facility (scale 1-5)*					2.1
CONTEXT-RELATED DETERMINANT LEVEL PRACTICE					
	Practice size (SD)		Offset variable		2545 (525)
% Privately insured (SD)					35.4 (11.2)
% Older than 65 years (SD)					14.4 (6.8)
Number of GPs working in computerised practice					206
Number of GPs using medical module information system					146
	Practice location:	Urban	Reference		108
		Semi-urban	-0.0195	0.1022	57
		Rural	-0.0804	0.1132	56
	Practice type:	Single-person	Reference		103
		Two-person	-0.0989	0.0954	41
		Group practice	-0.1641	0.1052	77

TABLE 3 INDIVIDUAL AND CONTEXT-RELATED DETERMINANTS OF THE NUMBER OF TESTS ORDERED BY 221 GPS IN 1997. (CONTINUED)

STUDY POPULATION CHARACTERISTICS	ELIGIBLE FOR INITIAL MULTILEVEL MODEL	B	SE B	TOTAL
% of GPs using desk top testing always for Hb, ESR and glucose				12.8
	Distance to imaging facility in km	0.0004	0.0087	6.2 (5.3)
	Distance to laboratory facility in km	0.0120	0.0130	2.3 (2.5)
LOCAL GP GROUP (N = 40)				
	Number of local GP groups receiving feedback on test ordering	-0.0678	0.2157	22
	Number of local GP groups making group plans for change	-0.0508	0.0994	26
	At least one GP in the GP group is involved in developing guidelines	-0.1220	0.1033	12
REGION (N = 5)				
	Number of diagnostic centres providing feedback on test ordering	-0.4776	0.1251	1

Abbreviations: β = Regression coefficient;
SE = Standard error;
SD = Standard deviation;
MRI = Magnetic Resonance Imaging

The second column shows the 11 determinants eligible for the initial multilevel model analysis, including practice size (offset variable).

* Personal opinions of GPs on test ordering 1= disagree....5= agree

Our study population differed from the total population of GPs in the Netherlands in some features, but we do not think that these differences influenced our results. Further, in the Netherlands diagnostic facilities only perform tests, when a physician orders them. Sometimes, however, diagnostic centres perform test cascades, depending on the results of the previous test. Further, only data from the diagnostic facility were available, so the tests that were ordered but not performed, e.g. because the patient did not visit the diagnostic centre, were not included. However, both situations probably constitute a small part of the ordered tests.

Based on the present results, it is tempting to recommend the introduction of problem-oriented forms in diagnostic facilities for GPs, however further study to replicate our findings is necessary. The problem-oriented form was developed as a quality improvement instrument, aimed at efficient and cost-efficient use of tests. Of course, it is also important to study patient-related factors, such as whether patients are actively demanding tests, and how to 'sell' such a cost-conscious approach to such demanding patients? These patient factors should be discussed with colleagues, as probably some of them may

have developed effective strategies for dealing with them. Despite the small influence of the local GP group in our study, many GPs mentioned social influence of colleagues as an important determinant of test ordering.

The medical coordinators of the diagnostic centres, who provide the feedback on test ordering and may as such be regarded as experts on this topic, could function as opinion leaders in these discussions. [29-31]

Based on the strong correlations we found between several factors and test ordering patterns, we conclude that a quality improvement programme, consisting of discussions on guidelines and feedback reports in a local GP group, and collaborating with a diagnostic centre, that uses problem-oriented test ordering forms and provides the feedback, appears to be a promising intervention to decrease overuse of GPs' test ordering.

TABLE 4 RESULTS OF FINAL TWO-LEVEL MULTIVARIABLE ANALYSIS

LEVEL	DETERMINANT		DIFFERENCE (%)	P	95% CI
Professional	Working full time or part time	5	(0) reference		
		4.5 days	-13.5	.210	-31.0; 8.5
		4 d	15.7	.204	-7.6; 45.0
		1.5-3.5 d	-14.3	.105	-28.9; 3.3
	Actively involved in developing guideline(s)	no	(0) reference		
		yes	-26.9	.013	-43.0; -6.4
Context-related practice	Practice type	single-person	(0) reference		
		two-person	-5.9	.516	-21.8; 13.1
		group	-18.0	.022	-30.9; -2.8
	Distance to imaging facility (per 10 km)		-9.4	.168	-21.9; 3.2
	Distance to laboratory facility (per 10 km)		19.1	.142	-7.4; 43.5
Context-related regional	Diagnostic centre providing feedback	no	(0) reference		
		yes	24.1	.311	-18.2; 88.3
	Problem-oriented form > 1 yr.	no	(0) reference		
		yes	-41.0	.001	-57.2 ; -18.7

Differences are percentage point changes compared with a reference category.

$r^2 = 0.304$

References

1. Ayanian JZ, Berwick DM. Do physicians have a bias toward action? A classic study revisited. *Medical Decision Making* 1991;11:154-8.
2. Zaat JO, van Eijk JT. General practitioners' uncertainty, risk preference, and use of laboratory tests. *Med Care* 1992;30(9):846-54.
3. Leurquin P, Van Casteren V, De Maeseneer J. Use of blood tests in general practice: a collaborative study in eight European countries. Eurosentinel Study Group. *Br J Gen Pract* 1995;45:21-5.
4. Ferrier BM, Woodward CA, Cohen M, Goldsmith CH. Laboratory tests: which physicians order more? *Can Fam Physician* 1991;37:347-52.
5. Wong ET. Improving laboratory testing: can we get physicians to focus on outcome? *Clinical Chemistry* 1995;41:1241-7.
6. Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G. Medicine. Communicating statistical information. *Science* 2000;290:2261-2.
7. Kaag ME, Wijkel D, de Jong D. Primary health care replacing hospital care--the effect on quality of care. *Int J Qual Health Care* 1996;8:367-73.
8. Kristiansen IS, Hjortdahl P. The general practitioner and laboratory utilization: why does it vary? *Fam Pract* 1992;9:22-7.
9. Little P, Cantrell T, Roberts L, Chapman J, Langridge J, Pickering R. Why do GPs perform investigations?: The medical and social agendas in arranging back X-rays. *Fam Pract* 1998;15:264-5.
10. Miller WL, McDaniel RR, Jr., Crabtree BF, Stange KC. Practice jazz: understanding variation in family practices using complexity science. *J Fam Pract* 2001;50:872-8.
11. Mabeck CE, Kragstrup J. Is variation a quality in general practice? *Scand J Prim Health Care Suppl* 1993;1:32-5.
12. Goold SD, Hofer T, Zimmerman M, Hayward RA. Measuring physician attitudes toward cost, uncertainty, malpractice, and utilization review. *J Gen Intern Med* 1994;9:544-9.
13. Bugter Maessen AM, Winkens RA, Grol RP, Knottnerus JA, Kester AD, Beusmans GH, et al. Factors predicting differences among general practitioners in test ordering behaviour and in the response to feedback on test requests. *Fam Pract* 1996;13:254-8.
14. Malcolm L, Wright L, Seers M, Davies L, Guthrie J. Laboratory expenditure in Pegasus Medical Group: a comparison of high and low users of laboratory tests with academics. *N Z Med J* 2000;113:79-81.
15. Cherkin DC, Deyo RA, Wheeler K, Ciol MA. Physician variation in diagnostic testing for low back pain. Who you see is what you get. *Arthritis Rheum* 1994;37:15-22.
16. van Merode GG, Stroink AE, Maarse JA, Goldschmidt HM. Impact of insurance coverage type on laboratory test ordering behaviour of general practitioners. *World Hosp Health Serv* 2000;36:7-12.
17. Sherwood P, Lyburn I, Brown S, Ryder S. How are abnormal results for liver function tests dealt with in primary care? Audit of yield and impact. *BMJ* 2001;322:276-8.
18. Winkens R, Dinant GJ. Evidence base of clinical diagnosis: Rational, cost effective use of investigations in clinical practice. *BMJ* 2002;324:783.
19. Armitage P BG, Matthews JNS, editor. *Statistical methods in medical research*. 4 ed: Blackwell, 2002.
20. Mittman B, Tonesk X, Jacobson P. Implementing clinical practice guidelines: social influence strategies and practitioner behavior change. *Qual Rev Bull* 1992;18:413-22.
21. Gorton TA, Cranford CO, Golden WE, Walls RC, Pawelak JE. Primary care physicians' response to dissemination of practice guidelines. *Arch Fam Med* 1995;4:135-42.
22. Solberg LI, Brekke ML, Fazio CJ, Fowles J, Jacobsen DN, Kottke TE, et al. Lessons from experienced guideline implementers: attend to many factors and use multiple strategies. *Jt Comm J Qual Improv* 2000;26:171-88.
23. Grimshaw JM, Shirran L, Thomas R, Mowatt G, Fraser C, Bero L, et al. Changing provider behavior: an overview of systematic reviews of interventions. *Med Care* 2001;39(8 Suppl 2):II2-45.
24. Wyatt TD, Reilly PM, Morrow NC, Passmore CM. Short-lived effects of a formulary on anti-infective prescribing--the need for continuing peer review? *Fam Pract* 1992;9:461-5.
25. von Ferber L, Bausch J, Koster I, Schubert I, Ihle P. Pharmacotherapeutic circles. Results of an 18-month peer-review prescribing-improvement programme for

general practitioners. *Pharmacoeconomics* 1999;16:273-83.
26. Gill PS, Makela M, Vermeulen KM, Freemantle N, Ryan G, Bond C, et al. Changing doctor prescribing behaviour. *Pharm World Sci* 1999;21:158-67.
27. Zaat JOM, Van Eijk JT, Bonte HA. Laboratory test form design influence test ordering by general practitioners in the Netherlands. *Med Care* 1992;30:189-98.
28. van Walraven C, Goel V, Chan B. Effect of population-based interventions on laboratory utilization: a time-series analysis. *JAMA* 1998;280:2028-33.
29. Lomas J, Enkin M, Anderson GM, Hannah WJ, Vayda E, Singer J. Opinion leaders vs audit and feedback to implement practice guidelines. Delivery after previous cesarean section. *JAMA* 1991;265:2202-7.
30. Soumerai SB, McLaughlin TJ, Gurwitz JH, Guadagnoli E, Hauptman PJ, Borbas C, et al. Effect of local medical opinion leaders on quality of care for acute myocardial infarction: a randomized controlled trial. *Jama* 1998;279:1358-63.
31. Borbas C, Morris N, McLaughlin B, Asinger R, Gobel F, Lomas J, et al. The role of clinical opinion leaders in guideline implementation and quality improvement. *Chest* 2000;118(2 Suppl):24s-32s.

CHAPTER III

Interventions to improve the use of diagnostic tests.

Trudy van der Weijden
Wim HJM Verstappen
Michel Wensing
Gerben ter Riet
Richard PTM Grol

For EPOC: Effective Practice and Organisation of Care - Cochrane Collaboration

Abstract

Background

Many different approaches have been adopted to improve health care providers' use of diagnostic tests. The objective of this systematic literature review is to describe the different approaches and to assess the effectiveness of the strategies aimed at influencing test use.

Methods

We searched Medline (1966 to 1997), the Cochrane Collaboration Effective Practice and Organisation of Care trials register (searched 2001) and snowballed reference lists of relevant articles. Published (quasi-)RCTs, controlled before and after studies, and interrupted time series analyses of any type of intervention to influence the test ordering behaviour of any type of health care professional, using tests to diagnose or monitor patient complaints, were included. Assessment of trial quality and data extraction was executed by two independent reviewers.

Results

In total 98 studies with 118 comparison groups were included. Seventy-one studies with 86 comparisons described results on changing absolute rate of test use. Twenty-seven studies with 32 comparisons focused on improving appropriateness of test use. Overall, results are heterogeneous due to differences in type or intensity of the intervention, the setting, or methodological differences between studies, such as differences in measurement periods (during or after intervention) or in correction for baseline differences. Probably, different strategies are needed for modifying overuse of tests versus improving appropriateness of test ordering behaviour. It is not clear that single strategies have less impact versus multifaceted strategies, but it seems important to focus the intervention at both the professional and the context. Audit and feedback seems effective for both decreasing absolute test rate and improving appropriateness of test use. Reminders by computer aided decision support improve the appropriateness of test use. Outreach visits, patient-mediated interventions and small group quality improvement deserve more attention.

Conclusions

There is no rule of thump for the choice of the intervention in effectively influencing test ordering behaviour. Next to generally accepted rules such as tailoring the intervention to the barriers for change, the aim of the intervention (modify overuse or improve appropriateness) should be considered in. In addition to professional-oriented interventions it seems important to consider the use of interventions that focus on organisational factors.

Introduction

Diagnostic tests (i.e. diagnostic procedures other than those performed in usual doctor-patient consultations, such as physical examination) are an important aspect of medical care in many areas of clinical practice. As doctors do not want to miss important diagnoses, the rapid advances in diagnostic technology has led to increased use of diagnostic tests. Overuse of diagnostic tests is a realistic danger in health care. It represents a potential threat to patient health, as false-positive findings can lead to harm from invasive diagnostic interventions (e.g. colonoscopy used to follow-up on patients with false-positive haemoccult tests), unnecessary treatment, or anxiety and labelling effects. In addition, it is a waste of resources.
According to diagnostic decision making theories the decision to order a test should at least be based on the pretest chance of the patient having the disorder and the seriousness of the suspected disorder. Other important considerations include, the diagnostic value of the test, the consequences of the test result for further decision making such as therapy, and the risk or financial costs accompanying use of the test. In daily medical practice, diagnostic decision making may be biased by professional factors or by structural aspects of the practice environment (context-related factors)[1] Well-known professional-related determinants of test ordering behaviour are the risk taking attitude [2,3], bias towards action [4], or other aspects of personality[5], as well as routines.[6] They may also be found in the interaction of the professional with the direct environment such as pressure of peers through social influence [6,7]; or of patients. [8,9]

Examples of context-related determinants of test ordering behaviour are the differences in quantity of test ordering between countries[10], between regions[11], or between academic and non-academic hospitals[12]. Workload[13], availability of diagnostic facilities[14], the organisation of the test ordering procedure[2], the remuneration system[15] and its impact on supplier-induced demand [16,17], and, finally, financial incentives or regulatory sanctions, are all examples of structural aspects of the practice environment.
In view of this knowledge this review describes the different approaches that have been reported in influencing test ordering behaviour in rigorous designs, and attempts to investigate the effectiveness of all interventions to influence diagnostic test use. In an attempt to reduce heterogeneity of studies, test ordering as part of delivering preventive services in patients without clinical uncertainty, which might imply different beliefs, attitudes, reactions and judgements of the care provider were excluded. In this review we hypothesised that changing absolute rate of test use (most often reducing general overuse of diagnostic tests) and improving appropriateness of test use (most often by explicit guidelines for certain disease-defined patient categories) are different behaviours that need different strategies. We hypothesised that single-faceted strategies in general have less impact than multi-faceted strategies.[18] And we hypothesised that studies that evaluated strategies that were context-oriented interventions have more impact than exclusively professional-oriented interventions.

Methods

Inclusion criteria for studies: only randomised (RCTs), and quasi-randomised controlled trials (CCT) controlled before and after (CBA) studies, or interrupted time series (ITS) with at least three measurement points before and after the intervention were considered for this review. Studies on any health care professional responsible for patient care are included. This review was targeted at all diagnostic testing; laboratory tests, imaging techniques, and function tests. The scope of this review is restricted to the use of diagnostic tests that are requested to confirm or to exclude a diagnosis, or monitoring patients with disease, signs or symptoms. Studies about tests used in situations without clinical uncertainty, which generally attempt to enhance test use (such as screening or pre-operative tests), were not included. Any type of (professional-oriented, organisational, financial, or regulatory) intervention aimed at influencing the use of diagnostic tests was considered. Objective measure of quantity (absolute rate) or quality (appropriateness) of test ordering behaviour in daily practice had to be reported in the study.

Search strategy: Medline was searched from 1966 to August 1997. The following mesh terms were combined to define 'quality assurance': quality-assurance-health-care, quality control, physician's-practice-patterns, education-medical-continuing, guidelines, medical-audit, peer review, reminder-systems, physician-incentive-plans, feedback, health-services-research, algorithms, cost-control. A combination of mesh and free text terms was used to define 'test ordering behaviour': diagnosis/education-standards-utilization, diagnostic-tests-routine, laboratory near test$, laboratory near use$, laboratory near ordering, test$ near use$, test$ near ordering. In addition, the Cochrane Collaboration EPOC Register of Trials was searched until 2001. Finally, all reference lists of identified studies and reviews were checked for relevant articles. Each abstract of all retrieved citations was checked by at least two of the authors independently on the inclusion criteria (TvdW/MW/WV).

Studies were screened for inclusion and data were extracted independently by at least two of the authors (TvdW/MW/WV/GT), using a standardised form developed in collaboration with the Cochrane Effective Practice and Organisation of Care (EPOC) Group. Reporting results separately for two subgroups of studies optimised comparability between studies. The subgroups were studies characterised by changing absolute rate of test use and studies targeting on improving appropriateness of test use. Some studies report on the effects of more than one intervention, and therefore comparison groups were the unit of analysis in describing the effects of the interventions. If outcomes were reported on separate (subgroups of) tests instead on the total number of diagnostic tests, they were summarised by calculating the sum of separate numbers of tests. This could not be done for qualitative outcomes because the denominators of the proportions of performance that was according to the guidelines varied widely. If more than one measurement point was reported at follow-up the average of the results on the various measurement points was calculated. If available, both the immediate effects (measurement during the intervention period) and the lasting effects of the interventions (measurement during the follow-up period) were analysed.

Results are reported in descriptive tables. For each individual study the effect was translated into a rough outcome scale on the difference in relative change between groups:

- \- the difference in relative change between the intervention versus the control group was in the opposite direction than expected/desired;
- 0 the difference in relative change between the intervention versus the control group was between –2% and +2%;
- \+ the difference in relative change between the intervention versus the control group was between +2% and 10%;
- ++ the difference in relative change between the intervention versus the control group was between +11% and 20%;
- +++ the difference in relative change between the intervention versus the control group was between higher than 20%.

Example: **Table 2**, first study (Eisenberg 1977): Relative change in number of tests ordered per admission in intervention group: 717 – 830 = -113 divided by 830 = -14%. Relative change in control group: 905 – 900 = +5 divided by 900 = +1%. The difference in relative change between groups is -14% -1% = -15%

In case no comparable baseline data could be extracted from the reported results the relative difference between intervention versus the control group was calculated.

Results

Characteristics of included studies.

Strictly applying the inclusion criteria generated 98 studies (with 118 comparisons) for inclusion in this review. All included studies reported on test ordering by physicians. **Table 1** shows the distribution of the type of studies along some crude criteria. The details of the studies are reported in **Tables 2-6.**

Most trials were conducted in the United States of America (n=60), 11 in the United Kingdom, three in Canada, 18 in continental Europe including Ireland, and six in Austral-Asia (Australia, New Zealand, Korea, Thailand, Bangladesh). The earliest trial was published in 1975. In 41 studies the practice setting was inpatients, 49 studies took place in outpatient care (both family medicine and outpatient clinics), in six studies it was mixed (Gama's study was executed both in the in and outpatient setting, and therefore two outcomes were reported: number of test per admission, and number of tests per patient-visit), and the setting was unclear in the remaining studies.

Seventy-one studies with 86 comparisons focused on changing absolute rate of test use, the "modify overuse" group. Twenty-seven studies with 39 comparisons targeted the improvement of appropriateness of test use, the "improve quality" group. In the "modify overuse" group the interventions were focused on one or a few disease-specific tests only in 15% of the studies (n=11), whereas this was 67% (n=18) for the "improve quality" group of studies. In the first group the authors reported explicit guidelines underlying the desired test ordering behaviour for 32% of these studies, whereas this was 88% in the second group of studies.

The number of published trials increased throughout the years (until 1980: n=11, '81 – '90: n= 38, '91 – '00: n=49). The aim of the studies and the type of strategies also changed throughout the years. The proportion of studies aiming at modifying overuse of tests increased through the years (until 1980: 55%, '81 – '90: 70%, '91 – '00: 76%). The proportion of studies evaluating multi-faceted strategy at least in one arm increased only in the eighties (until 1980: 0%, '81 – '90: 24%, '91 – '00: 14%). The proportion of studies evaluating context-oriented strategies increased in the nineties (until 1980: 27%, '81 – '90: 24%, '91 – '00: 49%).

There is some risk of methodological bias in all of the included trials. Over half of the studies (n=56) were randomised controlled trials. In 16 of these trials, we were confident that randomisation was properly executed at central level; in the remaining 39 trials the randomisation procedure was not clearly described. The allocation procedure was clearly concealed in 9 studies, clearly not concealed in 38 studies, and this criterion was scored as unclear in the remaining 51 studies. In 44 studies it was unlikely that the control group received the intervention, in the other 54 studies it was either unclear or likely that the control group received the intervention. Outcomes were assessed blindly in 18 studies, in 45 studies this was not the case, and in the remaining 35 studies this criterion was not clear. The number of professionals participating in the studies varied from 2 to 1483, but was not given in half of the studies (n=49) studies. Information on dropouts was also sparse. In 35% of the studies there was disagreement between the unit of analysis and the unit of randomisation. The clustering by physician was most often not taken into account in the analysis; therefore the results of the studies should be interpreted with caution because of bias towards effect.

Most studies compared the effect of the intervention with that of a control group without any intervention (usual care). In 6 studies (10 comparisons) the intervention was compared with another intervention. For reasons of comprehensiveness studies with multifaceted interventions (combinations of different type of interventions) were not described in a separate table. Nearly all interventions described educational materials or meetings as a component of the intervention. In this review, educational materials or meetings were regarded upon as a logical or necessary condition for an intervention to influence test ordering behaviour, not as a separate component of a multifaceted intervention. The category of organisational interventions typically shows a high rate of combinations with professional-oriented interventions.

Effects of strategies

There was large variation in the duration of the interventions. Interventions varied from two weeks to as long as 9 years (median 8 months, interquartile range 3-12 months). Most comparisons, namely 79, were on professional-oriented interventions, and 39 on context-oriented interventions (**Table 1**). Overall, a quarter (26%) of the interventions were aimed at 'improving quality'; 29% (23 out of 79) of the comparisons evaluating professional-oriented interventions, and 21% (8 out of 39) of comparisons evaluating context-oriented interventions respectively. The intervention types audit and feedback, reminders, and structural organisational interventions were evaluated most often.

PROFESSIONAL-ORIENTED INTERVENTIONS

*Distribution of educational materials/educational meetings (**Table 2A +2B**)*

For both type of studies on influencing test ordering behaviour, 'modify overuse' group and 'improve quality' group, the effects were small to moderate. Eisenberg's, Davidoff's and Stross' 1980 study showed a relevant decline at follow-up, after the intervention had stopped. The number of participants of the educational intervention might explain the magnitude of the effect; the study had only a small number of participants, it is therefore likely that the attention given to the participants was intensive. Davidoff's and Stross '80 study were also characterised by a small number of participants.

*Audit and feedback (**Table 3A + 3B**)*

Generally speaking, a consistent positive effect is seen in the 'modify overuse' group without a clear trend towards a specific content of the feedback given. Strong effects are seen e.g. in Winkens' study published in 1996, for which the long duration of the intervention (9 years) is striking. A strong rebound effect was seen in Cohen's study. Reason given: "Simple cost feedback mechanisms will not by themselves assure reduction in test usage; it requires effort to prepare physicians to use these data". Reason given for the opposite effect in Wones' study: "Perhaps the medium, e.g. a respected teacher, is more important than the message". Audit and feedback and information transfer also shows a consistent, and somewhat stronger effect. Chassin's study focussed on one test only. Only two studies, on audit and feedback including information transfer, are reported for the 'improve quality' group, with a strong effect in Kroenke's study.

*Reminders (**Table 4A + 4B**)*

Varying effects are seen in the 'modify overuse' group. The effect of computer aided decision support (studies of Thomas, Tierney, and Holleman) is disappointing for reducing overuse of tests. The effects of reminders in the 'improve quality' group seem less varying than reminders aimed at modifying overuse and more encouraging, also for computer aided decision support.

*Other professional-oriented interventions (**Table 5A + 5B**)*

In the 'modify overuse' group two small studies on educational outreach visits show reasonable effects during intervention, but the effect does not last in one study.

The availability of the patient's depression score before consultation, but not after the consultation, reduces laboratory testing. This is the only example of a patient-mediated intervention. Although the number of studies on small group quality improvement is limited, positive effects are shown in the 'modify overuse' group, but less so in the 'improve quality' group.

CONTEXT-ORIENTED INTERVENTIONS (*Table 6A + 6B*)

In the 'modify overuse' group, both the professional-related organisational interventions and the structural organisational interventions showed a rather consistent picture of positive results. The professional-related organisational interventions were most often characterised by demanding justification for test ordering by changing the organisation in such a way that an attending physician or a team was given a sort of supervising role. The structural organisational interventions typically

showed a more steering character, e.g. by applying strict protocols, or by changing the procedure of test ordering, or shifting responsibilities care setting. The three studies on financial interventions did not seem to lead to the desired effect, but the combined financial and organisational interventions seem more effective. Although numbers of studies are small.

TABLE 1 SUMMARY OF MAIN CHARACTERISTICS OF THE 118 COMPARISON GROUPS

		A: studies aimed at 'modify overuse', B: studies aimed at 'improve quality'				
TYPE OF STRATEGY	NUMBER OF STUDIES	A/B	NO OF COMPARISONS	RCT DESIGN	EXPLICIT GUIDELINES	MULTI-FACETED
PROFESSIONAL-ORIENTED INTERVENTIONS						
educational strategies	13 see Table 2	A	9	3 (33%)	2 (22%)	0 (0%)
		B	6	5 (83%)	4 (67%)	0 (0%)
audit and feedback	24 see Table 3	A	27	11 (41%)	8 (30%)	0 (0%)
		B	2	0 (0%)	2 (100%)	0 (0%)
reminders	22 see Table 4	A	11	8 (73%)	2 (18%)	1 (9%)
		B	12	8 (75%)	8 (67%)	1 (8%)
other	12 see Table 5	A	8	4 (50%)	2 (25%)	2 (25%)
		B	4	3 (75%)	4 (100%)	2 (50%)
SUB TOTAL	71*		79	42 (53%)		6 (8%)
CONTEXT-ORIENTED INTERVENTIONS						
professional-related organisational strategies	n = 9 see Table 6	A	8	4 (50%)	3 (38%)	1
		B	1	1 (100%)	1 (100%)	1
structural organisational strategies	n = 21see Table 6	A	17	7 (41%)	4 (24%)	0
		B	7	3 (43%)	4 (57%)	6
financial strategies	n = 3 see Table 6	A	3	2 (66%)	0 (0%)	0
		B	0	-	- -	-
combined organisational and financial strategies	n = 3 see Table 6	A	3	0 (0%)	0 (0%)	3
		B	0	-	- -	-
SUB TOTAL	36*		39	17 (43%)		
			118			

* some studies reported both on professional and context-oriented comparisons

Discussion

In this review we explored the variety and effectiveness of various interventions to influence test-ordering behaviour of physicians dealing with patients with signs or symptoms. An increasing number of studies have been executed, with modifying overuse of tests being the most common aim of the studies. Context-oriented strategies are increasingly evaluated in this set of studies. It was hypothesised that changing absolute rate of test use (most often reducing overuse of diagnostic tests) and improving appropriateness of test use (most often by following explicit guidelines) would require different interventions. The different findings for reminders (seem more effective for improving quality) and small group quality improvement (seem more effective for modifying overuse) seem to confirm this hypothesis for these type of strategies. No clear answer can be given on the hypothesis that multifaceted strategies are superior to single strategies. Interventions aimed at the contextual aspects of the practice environment seem to have more consistent effects than interventions aimed at direct professional-related issues such as attitude and knowledge exclusively. But, these context-oriented interventions are relatively more often multi-faceted. Although no clear conclusion can be drawn, facilitating the preferred diagnostic behaviour seems most potent through combined interventions aiming at both the professional and the context.

Overall, results are heterogeneous probably due to methodological differences between studies, such as differences in measurement periods (during or after intervention) or in correction for baseline differences. Distribution of educational materials and educational meetings should be looked upon as (necessary) parts of a multifaceted intervention. Audit and feedback, both with and without information transfer, show consistent and sometimes even strong effects on both changing the absolute rate of test use ("modifying overuse") as well as on improving appropriateness of test use ("improving quality"). There is no clear trend towards a specific content of the feedback given. Reminders show sometimes relevant but inconsistent effects on changing the absolute rate of test use. Reminders and computer aided decision support seem more suitable for improving the appropriateness of specific test use than for changing general overuse of tests. Small group quality improvement seems especially effective in changing the absolute rate of test use, perhaps because of the social influence that professionals can have on each other in applying this method. Context-oriented interventions show positive results, but there might be some bias; the type of designs is less rigorous (less RCTs) and they are quite often mixed with multifaceted strategies. Little is known about the effect of financial interventions, but the three studies were not promising in their results.

Many of the studies lacked power because we extracted only the data on test ordering behaviour whereas the power was calculated on broader outcomes. But in a literature review we look for trends in effects over categories of studies. A problem in data-extraction was the extent of reporting of the methods in the papers. Often very little information was given on how precisely the intervention was designed and executed.

It was also often not described if the intervention was based on insight on actual care and barriers for working according to guidelines.
A standardised format should be given in reporting about the intervention in these trials.
Recommendations based on this review should be considered in the context of the heterogeneity and methodological problems in the studies. The best choice among interventions to modify use of tests appears to be audit and feedback, small group quality improvement, or combinations of professional oriented interventions and organisational interventions. Appropriateness of test ordering can be improved by reminders, and combinations of professional oriented interventions and organisational interventions. Promising interventions, such as outreach visits, patient mediated interventions, small group quality improvement and combinations with organisational interventions, should be studied in well-designed randomised trials.

Acknowledgements
We thank Roberto Grilli, Jeremy Grimshaw, Cynthia Fraser for their useful comments.

Potential conflict of interest
None.

TABLE 2 EDUCATIONAL MATERIALS AND EDUCATIONAL MEETINGS

2A. Studies with the objective to change absolute test rate ('modify overuse')

STUDY	TYPE OF TESTS, (NO. OF TESTS)	INTERVENTION TYPE, DURATION (MONTHS).	DESIGN	ROFESS. PRE/ POST	OUTCOME	BASELINE I VERSUS C	FOLLOW-UP I VERSUS C	
Eisenberg 77	prothrombin time (1)	printed materials + lecture (1.5)	CBA	114/114	no tests/ admission	830 vs 900	717 vs 905	++
Schroeder 84	laboratory tests + radiology (?)	- printed + audiovisual materials + course (for medicals)	CBA			-	563 vs 592[1]	+
		- same programme for surgeons (12)		?/?	test costs/ physician/ year	-	380 vs 372[1]	0
Marton 85	outpatient laboratory utilisation (?)	printed materials inclusive list of charges (8)	RCT	57/?	no tests/ patient visit	1.61 vs 1.63	1.07 vs 1.34	++
Berwick 86	common blood test + X-rays (13)	printed materials + lecture (2)	CBA	35/35	no tests/ 1000 patient contacts/ physician	-	%-change: -012	+
Davidoff 89	little ticket tests (?)	printed materials + lecture (2)	RCT	24/24	no test/ admission	44.8 vs 43.4	32.0 vs 38.3[1]	++
Axt-Adam 93*	all (?)	- printed materials	CBA	507/507	no tests/ physician/ month	31.4 vs 31.1	34.9 vs 31.0[1]	-
	thyroid + kidney (2)	- printed materials + lecture (1)				31.0 vs 31.1	32.1 vs 31.0[1]	-
Oakeshott 94*	all X-ray (?)	printed materials (?)	RCT	62/62	no tests/ practice/ month	12.3 vs 15.3	8.1 vs 12.2[1]	++

2B. Studies with the objective to improve the appropriatness of test use ('improve quality')

Stross 80*	ESR, joint X-ray, latex test (3)	printed + audiovisual materials + lecture(12)	RCT	31/22	(tests done/indicated tests).100	34 vs 28%	51 vs 30%	++
Stross 83	X-chest, pulm. function, sputum (4)	printed+audiovisual materials+workshop for educational influentials (?)	RCT	?/?	(tests done/indicated tests).100	42 vs 47%-	45 vs 44%	+
White 85*	CPK-enzyme (1)	printed materials + lecture (<1)	RCT	?/103	(tests done/indicated tests).100	92 vs 72%	98 vs 72%	+
Bearcroft 94*	X-chest (1)	printed materials (1)	RCT#	210/7	(tests done/indicated tests).100	-	94 vs 92%[1]	+
Larsson 99	laboratory tests (14)	printed material + lecture (?)	CBA	63/63	ratio's meant to increase ratio's meant to decrease	-	desired change: 11 out of 14 vs 6 out of 14	+
STUDIES WITH AN INTERVENTION IN THE CONTROL GROUP								
Wirtschafter 86*	blood glucose +gases, X-chest (3)	printed materials + lecture versus printed materials	RCT	?/?	(tests done/indicated tests).100	-	31 vs 38%	-

* paper reports that explicit guidelines were available

high quality randomisation: clear description of central randomisation

[1]p-value < .05

follow-up measurement not after, but during the intervention period

The last column gives a standardised outcome for each individual study on the difference in relative change between intervention and control group

TABLE 3 CHARACTERISTICS AND RESULTS OF STUDIES WITH PROFESSIONAL-ORIENTED INTERVENTIONS / AUDIT AND FEEDBACK

3A. Studies with the objective to change absolute test rate ('modify overuse')

STUDY	TYPE OF TESTS, (NO. OF TESTS)	INTERVENTION TYPE, DURATION (MONTHS),	DESIGN	ROFESS. PRE/ POST	OUTCOME	BASELINE I VERSUS C	FOLLOW-UP I VERSUS C	
AUDIT AND FEEDBACK								
Eisenberg 77*	LDH + Ca (2)	On quantity (1)	CBA	?/?	(non-indicated tests/ tests done).100	51 vs 60	65 vs 77	-
Forrest 81	laboratory tests + radiology (?)	On costs (1.5)	CBA	?/?	test costs/ ward/ day	29.2 vs 28.0	28.4 vs 28.1	0
Cohen 82	total lab + imaging tests (?)	On costs (1)	RCT	?/?	no of tests/ admission	20.7 vs 26.4	20.5 vs 13.0[1]	-
Sherman 84	ECG holter monitoring (1)	Informed consent only (1)	CBA	?/?	no of tests/ hospital/ month	27 vs 42	32 vs 91	++
Marton 85	outpatient lab. utilisation (?)	On costs (8)	RCT	57/?	no of tests/patient contact	1.49 vs 1.63	1.04 vs 1.34	++
Berwick 86	common blood tests + X-rays (13)	- on costs of tests	CBA	35/35	no of test/ 1000 patients	-	-15.2%[1]	++
		- on yield of tests (2)				-	3.1% change	+
Tierney 87	blood/urine, ECG, X-chest/abd. (8)	On test results (4)	RCT	111/76	no of tests/ patient contact	0.61 vs 0.63	0.51 vs 0.56	+
Wones 87	commonly ordered tests (25)	- on quantity + costs	RCT	?/21	no tests/ patient/ day	-	3.27 vs 2.89	-
		- same feedback + group data for comparison (9)				-	3.10 vs 2.89	-
Pugh 89	diagnostic studies (?)	On daily costs (8)	CBA#	?/84	mean test costs	-	1488 vs 1592	+
Winkens 92*	laboratory tests (46)	On test quality (60)	CBA	85/?	no tests/ group physicians/ year	66250 vs 68750	50200 vs 73500[1]	+++
Gama 92*	haematology + clin. chemistry (?)	On quantity + costs (12)	CBA#	5/5	no tests/ admission	8.4 vs 8.1	8.8 vs 10.4	+++
					no tests/ patient contact	5.1 vs 1.4	4.0 vs 1.4	+++
Winkens 95*	X-ray, ECG, ultrasound (13)	On quantity and quality (30)	RCT	79/?	no tests/ physician/ year	110 vs 125	105 vs 142[1]	++
MacGowan '96	microbiological testing (?)	On quantity and yield of tests (24)	CBA#	?/?	no tests/ group physicians/ year	15596 vs 12806	14880 vs 14484	++
Winkens 96*	common tests (44)	On quantity and quality (108)	CBA	?/?	no tests/ group physicians/ year	114747 vs 191698	63062 vs 276401	+++

TABLE 3 CHARACTERISTICS AND RESULTS OF STUDIES WITH PROFESSIONAL-ORIENTED INTERVENTIONS / AUDIT AND FEEDBACK

3A. Studies with the objective to change absolute test rate ('modify overuse') CONTINUED

STUDY	TYPE OF TESTS, (NO. OF TESTS)	INTERVENTION TYPE, DURATION (MONTHS),	DESIGN	PROFESS. PRE/ POST	OUTCOME	BASELINE I VERSUS C	FOLLOW-UP I VERSUS C	
AUDIT AND FEEDBACK + INFORMATION TRANSFER								
Marton 85	outpatient lab. utilisation (?)	On costs + printed materials (8)	RCT	57/?	no tests/ patient contact	1.31 vs 1.63	1.03 vs 1.34	+
Chassin 86	X-pelvis (1)	On quantity + printed materials + lecture (4)	RCT	1483/?	no tests/ 1000 patients	73.4 vs 76.8	10.6 vs 36.4[1]	+++
Fowkes 86*	X-chest (1)	On quantity + printed materials (12)	CBA	?/?	no tests/ 1000 patients	326 vs 229	223 vs 199	++
Billi 87	laboratory tests + radiology (?)	On hospital charges + printed materials (12)	RCT#	132/132	test costs/ admission	-	119 vs 168	+++
Ruangkan. 93*	laboratory tests (?)	- on quantity + educational course	RCT	36/?	no tests/ admission	3.43 vs 3.44	3.63 vs 3.33	-
		- giving feedback to others- on quantity				3.34 vs 3.83	3.37 vs 3.21	-
		+ educat. course +giving feedback to others (6)				3.32 vs 3.27	3.36 vs 3.44	+
Freeborn 97	imaging tests lumbar spine (..)	On quantity + printed materials + lecture	CBA	95/95	no tests/ 1000 patient contacts	11.5 vs 11.6	11.9 vs 10.4	-
						16.1 vs 12.5	14.7 vs 10.2	-
Barwitz 99	lab, X-chest (..)	self-audit + printed materials + guideline development (12)	RCT	1/1	no tests/ patient contact	0.39 vs 0.40	0.24 vs 0.41[1]	+++
Kerry 00	imaging tests	On quantity + printed materials	RCT	175/175	no tests/ group physicians/ year	11960 vs 10300	11025 vs 10493	++
STUDIES WITH AN INTERVENTION IN THE CONTROL GROUP								
Schectman 91*	thyroid function (1)	feedback on quantity/quality + printed materials (2) versus printed material	CBA#	30/?	(tests done/indicated tests).100	53 vs 49%	64 vs 81%	-

3B. Studies with the objective to improve the appropriateness of test use ('improve quality')

Kroenke[85] '87*	sputum, urine cult., urinalysis (3)	on quantity and quality + lecture (2.5)	CBA	?/?	(indicated tests/tests done).100	45 vs 43%	65 vs 40%	+++
Oosterhuis[86] '95*	9 common indications (?)	on quality + printed materials (8)	CBA	78/28	(indicated tests/tests done).100	-	35 vs 26%	+

* paper reports that explicit guidelines were available

unit of analysis unequals unit of allocation

[1] p-value < .05

follow-up measurement not after, but during the intervention period

The last column gives a standardised outcome for each individual study on the difference in relative change between intervention and control group

TABLE 4 REMINDERS

4A. Studies with the objective to change absolute test rate ('modify overuse')

STUDY	TYPE OF TESTS, (NO. OF TESTS)	INTERVENTION TYPE, DURATION (MONTHS).	DESIGN	ROFESS. PRE/ POST	OUTCOME	BASELINE I VERSUS C	FOLLOW-UP I VERSUS C	
REMINDERS WITHOUT COMPUTERISATION								
Wexler 75	all (7)	Implicit reminder; DD print out (12)	RCT	?/?	no tests/ admission	-	13.6 vs 13.7	0
					no unnecesary tests/ admiss.	-	1.3 vs 2.3	+++
Wilson 82	laboratory tests + radiology (7)	Immediate access to computerised medical record (?)	RCT#	?/182	no tests/ admission	-9.6[1]	-	+
Tierney 90	outpatient diagnostic tests (?)	Implicit reminder on test charges (6)	RCT	121/74	No. tests/ patient contact ???	1.81 vs 1.72	1.56 vs 1.82[1]	++
Williams 86*	11 serum tests + X-chest (12)	Inter visit reminders + feedback + pr. materials + lecture (6)	CBA	143/?	(indicated tests/tests done).100	38 vs 42%	47 vs 60%	-
REMINDERS WITH COMPUTERISATION								
Thomas 83	laboratory, X-ray, ECG (7)	Computer aided decision support (12)	RCT	?/?	test costs/ patient/ year	-	101.4 vs 92.3	-
Tierney 88*	blood, ECG, urine, X-chest (8)	Comp. aided decision support; prediction of abnormal test result (6)	RCT	112/?	test costs/ patient contact	-	11.2 vs 12.3[1]	+
Holleman 96	all (7)	Computer aided decision support (3)	CBA	?/?	no tests/ patient contact	-	1.8 vs 1.8	0
Bates 97	laboratory + imaging tests (..)	Computer aided decision support, display of test charges (4)	RCT#	?/?	no tests/ admission	55.9 vs ...	48.4 vs 51.1	+
Harpole 97	abdomen X-rays (..)	Computer aided decision support + printed materials (4)	ITS	236/236	% cancelled tests	3% vs ...	4% vs ... [1]	???
Bates 99	laboratory tests (..)	Computer aided decision support about redundant tests (4)	RCT#	?/?	% cancelled tests	-	51 vs 27%[1]	+++
STUDIES WITH AN INTERVENTION IN THE CONTROL GROUP								
Pollack 91	lab. + radiology (?)	implicit reminders on patient's survival probability + audio-visual materials (?) versus audio-visual materials	RCT	?/94	no tests/ patient/ day	-	37.1 vs 33.2	-

* paper reports that explicit guidelines were available

\# high quality randomisation: clear description of central randomisation

p-value < 05

follow-up measurement not after, but during the intervention period

The last column gives a standardised outcome for each individual study on the difference in relative change between intervention and control group

TABLE 4 REMINDERS

4B. Studies with the objective to improve the appropriateness of test use ('improve quality')

STUDY	TYPE OF TESTS, (NO. OF TESTS)	INTERVENTION TYPE, DURATION (MONTHS).	DESIGN	ROFESS. PRE/ POST	OUTCOME	BASELINE I VERSUS C	FOLLOW-UP I VERSUS C	
REMINDERS WITHOUT COMPUTERISATION								
Bulpitt 76	urea, electrolytes (2)	concurrent report (12)	RCT	?/?	no tests/ patient/ year	-	1.5 vs 1.5	0
McDonald 80	tests ordered (?)	- concurrent report	RCT	31/31	(tests done/indicated tests).100	-	37 vs 15%	+++
		- concurrent report + printed materials (2.5)				-	37 vs 15%	+++
White 84*	ECG, serum potassium +digoxin (3)	concurrent report (3)	RCT#	?/?	no tests/ 1000 patients/ year	-	Ecg: 36 vs 29	++
							Pot: 117 vs 89†	+++
							Dig: 48 vs 17†	+++
Winickoff 85*	Hb creat pot chol urine ECG X-chest (7)	inter visit reminders + feedback on quality of tests (12)	RCT	?/?	(tests done/indicated tests).100	86 vs 84%	87 vs 87%	-
Stiell 94*	X-ankle, X-foot (2)	concurrent report + printed +audiovisual mat. + lecture (5)	CBA	?/?	no tests/ admission	1.14 vs 1.21	0.87 vs 1.27	+++
Auleley 97*	X-ankle, X-foot (2)	concurrent report + printed +audiovisual mat. + lecture (5)	RCT	91/?	(patients tested/all patients).100	98 vs 99%	79 vs 99%†	+++
REMINDERS WITH COMPUTERISATION								
McDonald 76*	mixed blood tests (>30)	computer aided decision support (4)	CBA	9	(tests done/indicated tests).100	-	61 vs 22%†	+++
McDonald 76	renal/liver funct, electr., Hb/Ht (7)	computer aided decision support (8)	RCT#	10/10	(indicated tests/tests done).100	-	36 vs 11%†	+++
Rogers 82*	renal function , pyelogram (5)	computer aided decision support (24)	RCT	?/?	(tests done/indicated tests).100	-	51 vs 40%	++
Overhage 97	laboratory (..)	computer aided decision support (7)	RCT#	86/86	(tests done/indicated tests).100		46 vs 22%	+++
STUDIES WITH AN INTERVENTION IN THE CONTROL GROUP								
Mazzuca 90*	glycolysed.Hb, fasting blood sugar, home-monitoring gluc. (3)	computer aided decision support+ pr. materials+lecture versus printed materials + lecture	CBA	114/?	(indicated tests/tests done).100	-	24 vs 21%	+

* paper reports that explicit guidelines were available

high quality randomisation: clear description of central randomisation

† p-value < .05

follow-up measurement not after, but during the intervention period

The last column gives a standardised outcome for each individual study on the difference in relative change between intervention and control group

TABLE 5 OUTREACH VISITS, PATIENT MEDIATED INTERVENTIONS, AND SMALL GROUP QUALITY IMPROVEMENT

5A. Studies with the objective to change absolute test rate ('modify overuse')

STUDY	TYPE OF TESTS, (NO. OF TESTS)	INTERVENTION TYPE, DURATION (MONTHS).	DESIGN	PROFESS. PRE/ POST	OUTCOME	BASELINE I VERSUS C	FOLLOW-UP I VERSUS C	
EDUCATIONAL OUTREACH VISITS								
Everett 85	laboratory tests (7)	Academic detailing (5)	RCT	16/7	no tests/ patient visit	-	4.8 vs 4.3	-
Everett 83	laboratory tests (7)	individual instruction and feedback on quantity +costs (5)	RCT	30/24	no tests/ admission	-	102 vs 120[1]	++
PATIENT MEDIATED INTERVENTIONS								
Linn 82	laboratory tests (7)	- depression score available before consultation	RCT	?/?	no tests/ patient visit	-	3.3 vs 4.3	+++
		- depression score available after consultation (?)				-	5.8 vs 4.3[1]	-
SMALL GROUP QUALITY IMPROVEMENT								
Schroeder 84	laboratory tests + radiology (7)	Small group quality improvement +feedback (12)	CBA	?/?	test costs/ physician/ year	-	544 vs 592	+
Fowkes 86*	9 common indications (7).	Small group quality improvement + lecture (2.5)	CBA	?/?	no tests/ admission	6.4 vs 6.1	3.8 vs 4.8	++
Fowkes 86*	X-chest (1)	Small group quality improvement + lecture (12)	CBA	?/?	no tests/ 1000 patients	290 vs 229	196 vs 199	++
STUDIES WITH AN INTERVENTION IN THE CONTROL GROUP								
Martin 80	lab. + radiology (7)	- small group quality improvement + pr. materials + lecture versus printed material + lecture	RCT	24/7	no tests/ admission	107 vs 102	51 vs 78[1]	+++

5B. Studies with the objective to improve the appropriateness of test use ('improve quality')

SMALL GROUP QUALITY IMPROVEMENT								
Palmer 85*	Ht, glucose, urine (3)	small group quality improvement + feedback (9)	RCT	548/7	(tests done/indicated tests).100	69 vs 67%	69 vs 68%	0
Gullion 88*	K, Ca, uric acid, glucose (4)	small group quality improvement + feedback on quantity/quality (2)	RCT	111/106	(tests done/indicated tests).100	59 vs 57%	50 vs 49%	-
Jones 93*	gastric endoscopy + X-ray (2)	small gr. quality impr. on consensus between GPs and specialists (4)	RCT#	179/179	no tests/ physician/ year	10.3 vs 8.8	9.7 vs 8.3	0
STUDIES WITH AN INTERVENTION IN THE CONTROL GROUP								
Hartmann 95*	gluc, HbA1c, alb, creat, lipids (7)	small group quality improv. (9) versus feedback on quality	CBA	17/17	(tests done/indicated tests).100	39 vs 49%	46 vs 44%	++

* paper reports that explicit guidelines were available

high quality randomisation: clear description of central randomisation

[1] p-value < .05

follow-up measurement not after, but during the intervention period

The last column gives a standardised outcome for each individual study on the difference in relative change between intervention and control group

TABLE 6 CONTEXT ORIENTED INTERVENTIONS

6A. Studies with the objective to change absolute test rate ('modify overuse')

STUDY	TYPE OF TESTS, (NO. OF TESTS)	INTERVENTION TYPE, DURATION (MONTHS).	DESIGN	PROFESS. PRE/ POST	OUTCOME	BASELINE I VERSUS C	FOLLOW-UP I VERSUS C	
PROFESSIONAL-RELATED ORGANISATIONAL INTERVENTIONS								
Marcy 81	non-routine tests on admission (?)	attending physician reviewing rationale of test orders (.5)	RCT	13/13	no tests/ admission	18.2 vs 16.4	16.6 vs 21.2[1]	++
Fowkes 86*	X-chest (1)	discussing justification of test orders + pr. materials (12)	CBA	?/?	no tests/ 1000 patients	290 vs 229	206 vs 199	++
Wachtel 86*	laboratory + X-rays + ECG (?)	attending physician reviewing rationale of test orders + small group quality improvement (3)	CBA	42/?	test costs/ admission	831 vs 695	580 vs 629	+++
Wachtel 90*	laboratory + X-rays + ECG (12)	local consensus devel. + demanding justificat. for orders (18)	CBA	161/?	test costs/ admission	534 vs 670	403 vs 554	+
Gottlieb 97	body imaging tests (?)	radiologist reviewing the rationale of test orders (2)	CBA	8/?	no tests/ 1000 patients	1190 vs 955	1127 vs 1204	+++
Naughton 94	all (?)	interdisciplinary geriatric team available (9)	RCT#	?/?	test costs/ admission	-	585 vs 897	+++
White 94	all (?)	interdisciplinary geriatric team available (6 beds) (?)	RCT#	?/?	no tests/ admission	-	4.4 vs 16.9[1]	+++
Koopmans 96	lab +imaging low back pain (..)	routine psychiatric consultation for low back pain patients	RCT#	4/4	no tests/ patient/ year (GPs)	-	1.16 vs 0.85	-
					(specialists)		1.16 vs 1.20	+
STRUCTURAL ORGANISATIONAL INTERVENTIONS								
Chambers 77	laboratory tests + radiology (?)	introduction of practice nurse with expanded role (12)	CBA	?/?	no tests/ 1000 patients	781 vs 1257	1349 vs 1837	-
Novich 85	PT/PTT, +common tests (15)	demanding written justification of orders + pr. materials (1)	CBA	?/?	no tests/ patient/ day	-	2.6 vs 3.5	+++
Fowkes 86*	X-chest (1)	change of test order form + printed materials(12)	CBA	?/?	no tests/ 1000 patients	245 vs 229	202 vs 199	+
Simmer 91	laboratory tests + radiology (?)	residents replaced by experienced staff (10)	RCT#	?/?	test costs/ admission	-	1315 vs 1649	++
Zaat 92*	all (?)	change of test order form + printed materials + educational course(7)	CBA	75/?	no tests/ 1000 patients	72.5 vs 75.8	59.3 vs 88.1[1]	+++
Gilio 93	all diagnostic services (?)	provision of desk top analysers (2)	RCT	26/19	no tests/ patient contact	1.8 vs 2.0	1.7 vs 1.6	-
Tierney 93*	all (?)	Comp. protocols, display former tests + charges (17)	RCT	?/?	test costs/ admission	-	1621 vs 1852[1]	++
Smithuis 94*	HDL+LDL, alk.fosf, total IgE (3)	change of test order form (6)	RCT	63/63	no tests/ patient contact	594 vs 566	182 vs 567[1]	+++
Ashworth 97	laboratory and imaging tests (..)	day care versus hospitalisation	CBA	?/?	test costs/ patient	-	12.3 vs13.4	+
		home care versus hospitilisation (11)					10.1 vs 13.4	+++
Murphy 96	blood and imaging tests (..)	GPs managing non-emergent patients at emergency department (13)	CBA	5/28	no tests/ patient contact	-	0.43 vs 0.65[1]	+++
Etter 97	laboratory and imaging tests (..)	managed care versus care without controlling access (12)	CBA	?/?	test costs/ patient/ year	135 vs 165	96 vs 178[1]	+++
Dahler-Eriksen 99	blood tests (..)	introduction of near patient CRP-testing (4)	CBA	64/?	no tests/ 1000 patients/ month	-	31.8 vs 33.7[1]	+

TABLE 6 CONTEXT ORIENTED INTERVENTIONS (CONTINUED)

6A. Studies with the objective to change absolute test rate ('modify overuse') (CONTINUED)

STUDY	TYPE OF TESTS, (NO. OF TESTS)	INTERVENTION TYPE, DURATION (MONTHS).	DESIGN	PROFESS PRE/ POST	OUTCOME	BASELINE I VERSUS C	FOLLOW-UP I VERSUS C	
STRUCTURAL ORGANISATIONAL INTERVENTIONS								
Board 00	laboratory (..)	re-engineered clinical pathway (12)	RCT	?/?	no tests/ patient contact	-	5.5 vs 8.3[1]	+++
Helgesen 00	lab, imaging;function tests (15)	introduction of specialised nurse versus urologists (?)	RCT#	?/?	no tests/ patient/ year	-	1.6 vs 2.3	+++
Lindley-Jones 00	imaging tests	introduction of triage X-ray requesting system nurse (2)	RCT#	?/?.	no tests/ 100 patients	-	68 vs 76[1]	++
					test yield/ 100 patients	-	54 vs 48[1]	+
Price 00	X-chest	change in order policy, no daily routine tests (26)	ITS	?/?	no tests/ patient/ day	1.0 SD 0.3	0.7 SD 0.2[1]	+++
FINANCIAL INTERVENTIONS								
Perkoff 76	laboratory tests + radiology (?)	Change of salary: from fee for service to prepaid practice (37)	RCT	?/?	no tests/ patient/ year	-	1.4 vs 0.7	-
Krasnik 90	all diagnostic services (?)	Change of sal.: to capitation based + mixed fee per item (12)	CBA	426/?	no tests/ 1000 patients	-	% change: 42.4[1]	-
STUDIES WITH AN INTERVENTION IN THE CONTROL GROUP								
Martin 80	lab. + radiology (?)	gift certificates if test ordering reduces + pr. materials + lecture (4) versus printed materials + lecture	RCT#	24/?	no tests/ admission	112 vs 102	100 vs 78	-
COMBINED ORGANISATIONAL AND FINANCIAL INTERVENTION								
Kerr 96	all (?)	change of test order form + budget holding + feedback on quantity/costs (9)	CBA	170/?	test costs/ patient visit	9.9 vs 9.3	6.6 vs 7.2	+
Wairaven 98	laboratory (11)	change test order from + change to top funding policy + printed materials (24)	ITS	?/?	no tests/ 1000 patients	.....	12 – 96% drop[1]	++
Makela 98	laboratory and imaging tests (..)	change in capitation + change of working patterns towards local population responsibility (48)	CBA	?/?	no tests/ patient/ year	3.8 vs 4.0	3.7 vs 3.6	-

* paper reports that explicit guidelines were available

high quality randomisation: clear description of central randomisation

[1] p-value <.05

follow-up measurement not after, but during the intervention period

The last column gives a standardised outcome for each individual study on the difference in relative change between intervention and control group

TABLE 6 CONTEXT ORIENTED INTERVENTIONS

6B. Studies with the objective to improve the appropriateness of test use ('improve quality')

STUDY	TYPE OF TESTS, (NO. OF TESTS)	INTERVENTION TYPE, DURATION (MONTHS).	DESIGN	PROFESS PRE/ POST	OUTCOME	BASELINE I VERSUS C	FOLLOW-UP I VERSUS C	
PROFESSIONAL-REATED ORGANISATIONAL INTERVENTIONS								
Jin 93*	X-chest+sputum (2)	intensive supervision + feedback on quantity + feedback on quality by patients (12)	RCT	?/?	(tests done/indicated tests).100	X-chest: sputum:	98 vs 80% 98 vs 70%	++
structural organisational interventions								
Bass 86*	urine, electrolytes, IVP (3)	expanded role of med. assistant + inter visit remind. + patient reminders (60)	RCT#	?/34	(tests done/indicated tests).100	-	63 vs 65%	0
Emslie 93*	3 lab. tests + semen analysis (4)	semen analysis packs available + concurrent report reminder + printed materials (8)	RCT	?/?	(tests done/indicated tests).100	-	58 vs 26% †	+++
Isouard 99	laboratory tests (..)	TQM + feedback on test use + change in test order form + printed materials (15)	CBA	?/?	(tests done/indicated tests).100	78 vs 82	88 vs 80	++
Saint 99	urine tests (2)	small group quality improvement + organisational change + printed materials (?)	CBA	?/?	(tests done/non-indicated tests).100	85 vs ..	64 vs 80†	++
STUDIES WITH AN INTERVENTION IN THE CONTROL GROUP								
Wirtschafter 86*	blood glucose +gases, X-chest (3)	individual instruction by toll-free telephone line + printed materials + lecture (8) versus printed materials	RCT	?/?	(tests done/indicated tests).100	-	45 vs 38%	+
Mazzuca 90*	glycolysed.Hb, fasting blood sugar, home-monitoring, glucose (3)	- computer aided decision support + printed materials + lecture + provision of desk top analysers, self care forms versus pr. materials + lecture + computer aided dec. support	CBA	114/?	(indicated tests/tests done).100	-	37 vs 24%†	++
		- *idem* + on call patient educator available (11) versus *idem* + analyser + forms				-	25 vs 37% †	-

* paper reports that explicit guidelines were available

high quality randomisation: clear description of central randomisation

†p-value < .05

follow-up measurement not after, but during the intervention period

The last column gives a standardised outcome for each individual study on the difference in relative change between intervention and control group

References of studies in the text

1. Weijden T van der, Bokhoven MA van, Dinant GJ, Hasselt CM Van, Grol RPTM. Understanding laboratory testing in diagnostic uncertainty: a qualitative study in general practice. *Br J Gen Pract* 2002;52:974-980.
2. Zaat JO, Eijk JT van. General practitioners' uncertainty, risk preference, and use of laboratory tests. *Med Care* 1992;30:846-54.
3. Epstein AM, Begg CB, McNeil BJ. The effects of physicians' training and personality on test ordering for ambulatory patients. *Am J Public Health* 1984;74:1271-3.
4. Ornstein SM, Markert GP, Johnson AH, Rust PF, Afrin LB. The effect of physician personality on laboratory test ordering for hypertensive patients. *Med Care* 1988;26:536-43.
5. Williams SV, Eisenberg JM, Pascale LA, Kitz DS. Physicians' perceptions about unnecessary diagnostic testing. *Inquiry* 1982;19:363-70.
6. Freeborn DK, Baer D, Greenlick MR, Bailey JW. Determinants of medical care utilization: physicians' use of laboratory services. *Am J Public Health* 1972;62:846-53.
7. Williams SV, Eisenberg JM, Pascale LA, Kitz DS. Physicians' perceptions about unnecessary diagnostic testing. *Inquiry* 1982;19:363-70.
8. Berwick DM, Weinstein MC. What do patients value? Willingness to pay for ultrasound in normal pregnancy. *Med Care* 1985;23:881-93.
9. Dolan JG, Bordley DR, Miller H. Diagnostic strategies in the management of acute upper gastrointestinal bleeding: patient and physician preferences. *J Gen Intern Med* 1993;8:525-9.
10. Leurquin P, Casteren V van, Maeseneer J de. Eurosentinel Study Group. Use of blood tests in general practice: a collaborative study in eight European countries. *Br J Gen Pract* 1995;45:21-5.
11. Verstappen W, Ter Riet G, Dubois WI et al Variation in test ordering behaviour of general practitioners: professional or context-related factors? *Fam Pract* 2004;21:385-93
12. Brook RH, Park RE, Chassin MR, Solomon DH, Keesey J, Kosecoff J. Predicting the appropriate use of carotid endarterectomy, upper gastrointestinal endoscopy angiography. *N Eng J Med* 1990;323:1173-7.
13. Daniels M, Schroeder SA. Variation among physicians in use of laboratory tests. II Relation to clinical productivity and outcomes of care. *Med Care* 1977;15:482-7.
14. Hillman BJ, Joseph CA, Marry MR, Sunshine JH, Kennedy SD, Noether M. Frequency and costs of diagnostic imaging in office practice. A comparison of self-referring and radiologist-referring physicians. *N Engl J Med* 1990;323:1604-8.
15. Shimmura K. Effects of different renumeration methods on general medical practice: a comparison of capitation and fee-for-service payment. *Int J Health Planning Management* 1988;3:254-8.
16. Iglehart JK. Congress moves to regulate self-referral and physicians' ownership of clinical laboratories. *N Eng J Med* 1990;322:1682-7.
17. Rice TH. The impact of changing medicare reimbursement rates on physician-induced demand. *Med Care* 1983;21:803-15.
18. Wensing M, Weijden T van der, Grol R. Implementing guidelines and innovations in general practice: which interventions are effective? *Br J Gen Pract* 1998;48:991-7.

References of included studies

- Auleley GR, Ravaud P, Giraudeau B, Kerboull L, Nizard R, Massin P, Garreau de Loubresse C, Vallee C, Durieux P. Implementation of the Ottawa ankle rules in France. A multicenter randomized controlled trial. *JAMA* 1997;277:1935-9.
- Ashworth A, Khanum S. Cost-effective treatment for severely malnourished children: what is the best approach? *Health Policy & Planning 1997*;12:115-21.
- Axt-Adam, van der Wouden, Hoek, van der Does. Het effect van nascholing op het aanvragen van laboratoriumdiagnostiek door huisartsen. *Huisarts Wet* 1993;36:451-4.
- Barwitz, HJK. Erkältung: eine Handlungsempfehlung. *Z Allg Med* 1999;75:932-38.
- Bass MJ, McWhinney IR, Donner A. Do family physicians need medical assistants to detect and manage hypertension? *Can Med Assoc J* 1986;134:1247 55.
- Bates DW, Kuperman GJ, Jha A, Teich JM, Orav EJ, Ma'luf N, Onderdonk A, et al. Does the computerized display of charges affect inpatient ancillary test utilization? *Arch Int Med* 1997;157:2501-8.

- Bates DW, Kuperman GJ, Rittenberg E, Teich JM, Fiskio J, Ma'luf N Onderdonk A Wybenga D, Winkelman J, Brennan TA, Komaroff AL, Tanasijevic M. A randomized trial of a computer-based intervention to reduce utilization of redundant laboratory tests. *Am J Med* 1999;106:144-50.
- Bearcroft PW, Small JH, Flower CD. Chest radiography guidelines for general practitioners: a practical approach. *Clin Radiol* 1994;49:56-8.
- Berwick DM, Coltin KL. Feedback reduces test use in a health maintenance organization. *JAMA* 1986;255:1450-4.
- Billi JE, Hejna GF, Wolf FM, Shapiro LR, Stross JK. The effects of a cost-education program on hospital charges. *J Gen Intern Med* 1987;2:306-11.
- Board N, Brennan N, Caplan G. Use of pathology services in re-engineered clinical pathways. *J Qual Clin Pract* 2000;20:24-29.
- Bulpitt CJ, Beilin LJ, Coles EC, Dollery CT, Johnson BF, Munro-Faure AD, Turner SC. Randomised controlled trial of computer-held medical records in hypertensive patients. *BMJ* 1976;1:677-9.
- Chambers LW, Bruce-Lockhart P, Black DP, Sampson E, Burke M. A controlled trial of the impact of the family practice nurse on volume, quality, and cost of rural health services. *Med Care* 1977;15:971-81.
- Chassin MR, McCue SM. A randomized trial of medical quality assurance. Improving physicians' use of pelvimetry. *JAMA* 1986;256:1012-6.
- Cohen D. Does cost information availability reduce physician test usage? A randomized clinical trial with unexpected findings. *Med Care* 1982;20:286-92.
- Dahler-Eriksen, BS, Lauritzen T, Lassen JF, Lund ED, Brandslund I. Near-patient test for C-Reactive protein in general practice: assessment of clinical, organizational, and economic outcomes. *Clin Chem* 1999;45:478-85.
- Davidoff F, Goodspeed R, Clive J. Changing test ordering behavior. A randomized controlled trial comparing probabilistic reasoning with cost-containment education. *Med Care* 1989;27:45-58.
- Eisenberg JM. An educational program to modify laboratory use by house staff. *J Med Education* 1977;52:578-81.
- Eisenberg JM, Williams SV, Garner L, Viale R, Smits H. Computer-based audit to detect and correct overutilisation of laboratory tests. *Med Care* 1977;15:915-21.
- Emslie C, Grimshaw J, Templeton A. Do clinical guidelines improve general practice management and referral of infertile couples. *BMJ* 1993;306:1728-31.
- Etter JF, Perneger TV. Health care expenditures after introduction of a gatekeeper and a global budget in a Swiss health insurance plan. *J Epidemiol Community Health* 1998;52:370-76.
- Everett GD. Impact of supervision by medical teachers and in-patient test control programmes on the out-patient test utilization of residents. *Med Educ* 1985;19:138-42.
- Everett GD, deBlois CS, Chang PF, Holets T. Effect of cost education, cost audits, and faculty chart review on the use of laboratory services. *Arch Intern Med* 1983;143:942-4.
- Forrest JB, Ritchie WP, Hudson M, Harlan JF. Cost containment through cost awareness. A strategy that failed. *Surgery* (St Louis) 1981;90:154-8.
- Fowkes FG, Hall R, Jones JH, Scanlon MF, Elder GH, Hobbs DR, Jacobs A, Cavill IA, Kay S. Trial of strategy for reducing the use of laboratory tests. *BMJ* 1986;292:883-5.
- Fowkes FG, Davies ER, Evans KT, Green G, Hartley G, Hugh AE, Nolan DJ, Power AL, Roberts CJ, Roylance J. Multicentre trial of four strategies to reduce use of a radiological test. *Lancet* 1986;1:367-70.
- Freeborn DK, Shye D, Mullooly JP, Eraker S, Romeo J. Primary care physicians' use of lumbar spine imaging tests: effects of guidelines and practice pattern feedback. *J Gen Int Med* 1997;12:619-25.
- Gama R, Nightingale PG, Broughton PMG, Peters M, Bradby GVH, Berg J, Ratcliffe JG.. Feedback of laboratory usage and cost data to clinicians: does it alter requesting behaviour? *Ann Clin Biochem* 1991;28:143-9.
- Gama R, Nightingale PG, Broughton PM, Peters M, Ratcliffe JG, Bradby GV, Berg J. Modifying the request behaviour of clinicians. *J Clin Pathol.* 1992; 45: 248-9.
- Gilio C, Buntinx F, de Kezel O, Scheys I. The influence of a desk-top analyser on the number of laboratory tests used in daily general practice. A randomized controlled trial. *Family Practice* 1993;10:118-23.
- Gottlieb RH, Hollenberg GM, Fultz PJ, Rubens DJ. Radiologic consultation: effect on inpatient diagnostic imaging evaluation in a teaching hospital. *Acad Radiol* 1997;4:217-21.
- Gullion DS, Tschann JM, Adamson TE, Coates TJ. Management of hypertension in private practices: a randomized controlled trial in continuing medical education. *J Continuing Educ Health Professions* 1988;8:239-55.
- Harpole LH, Khorasani R, Harpole LH, Khorasani R, Fiskio J, Kuperman GJ, Bates DW. Automated evidence-based critiquing of orders for abdominal

radiographs: impact on utilization and appropriateness. *J Am Med Informatics Association* 1997;4:511-21.
- Hartmann P, Bott U, Grüßer M, Kronsbein P, Jörgens V. Effects of peer-review groups on physicians' practice. *Eur J Gen Pract* 1995;1:107-12.
- Helgesen F, Andersson SO, Gustafsson O, Varenhorst E, Goben B, Carnock S, Sehlstedt L, Carlsson P, Holmberg L, Johansson JES. Follow-up of prostate cancer patients by on-demand contacts with a specialist nurse: a randomized study. *Scand J Urol Nephrol* 2000;34:55-61.
- Holleman DR Jr, Simel DL. Effectiveness of automatic diagnostic test result feedback on outpatient laboratory and radiology testing in veterans. A controlled trial. *Med Care* 1996;34:857-61.
- Isouard G. A quality management intervention to improve clinical laboratory use in acute myocardial infarction. *Med J Australia* 1999;170:11-4.
- Jin BW, Kim SC, Mori T, Shimao T. The impact of intensified supervisory activities on tuberculosis treatment. *Tubercle and Lung Disease* 1993;74:267-72.
- Jones RH, Lydeard S, Dunleavey J. Problems with implementing guidelines: a randomised controlled trial of consensus management of dyspepsia. *Quality in Health Care* 1993;2:217-21.
- Kerr D, Malcolm L, Schousboe J, Pimm F. Successful implementation of laboratory budget holding by Pegasus Medical Group. *N Z Med J* 1996;109:334-7.
- Kerry S, Oakeshott P, Dundas D, Williams J. Influence of postal distribution of The Royal College of Radiologists' guidelines, together with feedback on radiological referral rates, on X-ray referrals from general practice: a randomized controlled trial. *Fam Pract* 2000;17:46-52.
- Koopmans GT, Meeuwesen L, Huyse FJ, Heimans JJ. Effects of psychiatric consultation on medical consumption in medical outpatients with low back pain. *Gen Hospital Psychiatry* 1996;18:145-54.
- Krasnik A, Groenewegen PP, Pedersen PA, Scholten von P, Mooney G, Gottschau A, Flierman HA, Damsgaard MT. Changing remuneration systems: effects on activity in general practice. *BMJ* 1990;300:1698-701.
- Kroenke K, Hanley JF, Copley JB, Matthews JI, Davis CE, Foulks CJ, Carpenter JL. Improving house staff ordering of three common laboratory tests. Reductions in test ordering need not result in underutilization. *Med Care* 1987;25:928-35.
- Larsson, A, Biom S, Wernroth ML, Hulten G, Tryding N. Effects of an education programme to change clinical laboratory testing habits in primary care. *Scand J Prim Health Care* 1999;17:238-43.
- Lindley-Jones M, Finlayson BJ. Triage nurse requested X ray. Are they worthwhile? *J Accid Emerg Med* 2000;17:103-7.
- Linn LS, Yager J. Screening of depression in relationship to subsequent patient and physician behavior. *Med Care* 1982;20:1233-40.
- MacGowan AP, Feeney R, Brown I, Mcculloch S, Reeves D, Lovering A. Routine feedback to GPs who request microbiological tests is effective [letter; comment] *BMJ* 1996;312:1481.
- Mäkelä M, Sainio S, Åström M, Bergström M. Local population responsibility in Finnish health centres in 1989-1993. *Eur J Public Health* 1998;8:313-8.
- Marcy WL, Miller ST, Zwaag RV. Modification of admission diagnostic test ordering by residents. *J Fam Prac* 1981;12:141-2.
- Martin AR, Wolf MA, Thibodeau LA, Dzau V, Braunwald E. A trial of two strategies to modify the test-ordering behavior of medical residents. *N Engl J Med* 1980;303:1330-6.
- Marton KI, Tul V, Sox HC Jr. Modifying test-ordering behavior in the outpatient medical clinic. A controlled trial of two educational interventions. *Arch Intern Med* 1985;145:816-21.
- Mazzuca SA, Vinicor F, Einterz RM, Tierney WM, Norton JA, Kalasinski LA. Effects of the clinical environment on physicians' response to post-graduate medical education. *Am Educ Research* J 1990;27:473-88.
- McDonald CJ, Wilson GA, McCabe GP Jr. Physician response to computer reminders. *JAMA* 1980;244:1579-81.
- McDonald CJ. Protocol-based computer reminders, the quality of care and the non-perfectability of man. *N Engl J Med* 1976;295:1351-5.
- McDonald CJ. Use of computer to detect and respond to clinical events: Its effect on clinician behavior. *Ann Intern Med* 1976;84:162-7.
- Murphy AW, Bury G, Plunkett PK, Gibney D, Smith M, Mullan E, Johnson Z. Randomised controlled trial of general practitioner versus usual medical care in an urban accident and emergency department: process, outcome, and comparative cost. *BMJ* 1996;312:1135-42.
- Naughton BJ, Moran MB, Feinglass J, Falconer J, Williams ME. Reducing hospital costs for the geriatric patient admitted from the emergency department: a randomized trial. *J Am Geriatr Soc* 1994;42:1045-9.
- Novich M, Gillis L, Tauber AI. The laboratory test justified. An effective means

to reduce routine laboratory testing. *Am J Clin Pathol* 1985;84:756-9.
- Oakeshott P, Kerry SM, Williams JE. Randomized controlled trial of the effect of the Royal College of Radiologists' guidelines on general practitioners' referrals for radiographic examination. *Br J Gen Pract* 1994;44:197-200.
- Oosterhuis WP, Bosch WJHM van den, Calseijde JF van de, Veldhuis BRJ, Hoogen HJM van den, Kaathoven LGIM van, Kolnaar B, Meyers-Koopman L, Schuurmans MMJ. Ervaringen met verschillende methoden voor het verbeteren van het aanvragen van laboratoriumbepalingen in de huisartsenpraktijk. *Ned Tijdschr Klin Chem* 1995;20:72-5.
- Overhage JM, Tierney WM, Zhou XH, McDonald CJ. A randomized trial of "corollary orders" to prevent errors of omission. *J Am Med Informatics Association* 1997;4:364-75.
- Palmer RH, Louis TA, Hsu LN, Peterson HF, Rithrock JK, Strain R, Thompson MS, Wright EA. A randomized controlled trial of quality assurance in sixteen ambulatory care practices. *Med Care* 1985;23:751-70.
- Perkoff GT, Kahn L, Haas PJ. The effect of an experimental prepaid group practice on medical care utilization and cost. *Med Care* 1976;14:432-49.
- Pollack MM, Getson PR. Pediatric critical care cost containment: combined actuarial and clinical program. *Crit Care Med* 1991;19:12-20.
- Price MB, Chellis Grant MJ, Welkie K. Financial impact of elimination of routine chest radiographs in a pediatric intensive care unit. *Crit Care Med* 1999;27:1588-93.
- Pugh JA, Frazier LM, Delong E, Wallace AG, Ellenbogen P, Linfox SE. Effect of daily charge feedback on inpatient charges and physician knowledge and behavior. *Arch Intern Med* 1989;149:426-9.
- Rogers JL, Haring OM, Wortman PM, Watson RA, Goetz JP. Medical information systems: assessing impact in the areas of hypertension, obesity and renal disease. *Med Care* 1982;20:63-74.
- Rogers JL, Haring OM. The impact of a computerized medical record summary system on incidence and length of hospitalization. *Med Care* 1979;17:618-30.
- Ruangkanchanasetr S. Laboratory investigation utilization in pediatric out-patient department Ramathibodi Hospital. *J Med Assoc Thai* 1993;76 Suppl 2:194-208.
- Saint S, Scholes D, Fihn SD, Farrell RG, Stamm WE. The effectiveness of a clinical practice guideline for the management of presumed uncomplicated urinary tract infection in women. *Am J Med* 1999;106:636-41.
- Schectman JM, Elinsky EG, Pawlson LG. Effect of education and feedback on thyroid function testing strategies of primary care clinicians. *Arch Intern Med* 1991;151:2163-6.
- Schroeder SA, Myers LP, McPhee SJ, Showstack JA, Simborg DW, Chapman SA, Leong JK. The failure of physician education as a cost containment strategy. Report of a prospective controlled trial at a university hospital. *JAMA* 1984;252:225-30.
- Sherman H. Surveillance effects on community physician test ordering. *Med Care* 1984;22:80-3.
- Simmer TL, Nerenz DR, Rutt WM, Newcomb CS, Benfex DW. A randomized controlled trial of an attending staff service in general internal medicine. *Med Care* 1991;29:js31-js40.
- Smithuis LOMJ, Geldrop WJ van, Lucassen PLBJ. Beperking van het laboratoriumonderzoek door een probleemgeorienteerd aanvraagformulier. Een partiele implementatie van NHG-standaarden. *Huisarts Wet* 1994;37:464-6.
- Stiell IG, McKnight RD, Greenberg GH, McDowell I, Nair RC, Wells GA, Johns C, Worthington JR. Implementation of the Ottawa ankle rules. *JAMA*.1994;271:827-32.
- Stross J. Evaluation of a continuing education program in Rheumatoid arthritis. *Arthritis and Rheumatism* 1980;23:846-9.
- Stross JK, Hiss RG, Watts CM, Davis WK, MacDonald R. Continuing education in pulmonary disease for primary care physicians. *Am Rev Resp Dis* 1983;127:739-46.
- Stuart ME, Macuiba J, Heidrich F, Farrell RG, Braddick M, Etchison S. Successful implementation of an evidence-based clinical practice guideline: acute dysuria/urgency in adult women. *HMO Practice* 1997;11:150-7.
- Thomas JC, Moore A, Qualls PE. The effect on cost of medical care for patients treated with an automated clinical audit system. *J Med Syst* 1983;7:307-13.
- Tierney WM, McDonald CJ, Martin DK, Rogers MP. Computerized display of past results: effect on outpatient testing. *Ann Intern Med* 1987;107:569-74.
- Tierney WM, McDonald CJ, Hui SL, Martin DK. Computer predictions of abnormal test results effects on outpatient testing. *JAMA* 1988;259:1194-8.
- Tierney WM, Miller ME, McDonald CJ. The effect on test ordering of informing physicians of the charges for outpatient diagnostic tests. *N Engl J Med* 1990;322:1499-504.

- Tierney WM, Miller ME, Overhage JM, McDonald CJ. Physician inpatient order writing on microcomputer workstations. Effects on resource utilization. *JAMA* 1993;269:379-83.
- Wachtel T, Moulton AW, Pezzullo J, Hamolsky M. Inpatient management protocols to reduce health care costs. *Med Decis Making* 1986;6:101-9.
- Wachtel TJ, O'Sullivan P. Practice guidelines to reduce testing in the hospital. *J Gen Intern Med* 1990;5:335-41.
- Walraven C van, Goel V, Chan B. Effect of population-based interventions on laboratory utilization. A time-series analysis. *JAMA* 1998;280:2028-33.
- Wexler JR, Swender PT, Tunnessen WW Jr, Oski FA. Impact of a system of computer-assisted diagnosis. Initial evaluation of the hospitalized patient. *Am J Dis Child* 1975;129:203-5.
- White KS, Lindsay A, Pryor TA, Brown WF, Walsh K. Application of a computerized medical decision-making process to the problem of digoxin intoxication. *J Am Coll Cardiol* 1984;4:571-6.
- White CW, Albanese MA, Brown DD, Caplan RM. The effectiveness of continuing medical education in changing the behaviour of physicians caring for patients with acute myocardial infarction. *Ann Int Med* 1985;102:686-92.
- White SJ, Powers JS, Knight JR, Harrell D, Varnell L, Vaughn C, Brawner D, Burger MC. Effectiveness of an inpatient geriatric service in a university hospital. *J Tenn Med Assoc* 1994;87:425-8.
- Williams SV, Eisenberg JM. A controlled trial to decrease the unnecessary use of diagnostic tests. *J Gen Intern Med* 1986;1:8-13.
- Wilson GA, McDonald CJ, Mc Gabe McCabe GP Jr. The effect of immediate access to a computerized medical record on physician test ordering: a controlled clinical trial in the emergency room. *Am J Public Health* 1982;72:698-702.
- Winickoff RN, Wilner S, Neisuler R, Barnett GO. Limitations of provider interventions in hypertension quality assurance. *Am J Publ Health* 1985;75:43-6.
- Winkens RA, Pop P, Grol RP, Kester AD, Knottnerus JA. Effect of feedback on test ordering behaviour of general practitioners. *BMJ* 1992;304:1093-6.
- Winkens RA, Pop P, Bugter Maessen AM, Grol RP, Kester AD, Beusmans GH, Knottnerus JA. Randomised controlled trial of routine individual feedback to improve rationality and reduce numbers of test requests. *Lancet* 1995;345:498-502.
- Winkens RA, Ament AJ, Pop P, Reniers PH, Grol RP, Knottnerus JA. Routine individual feedback on requests for diagnostic tests: an economic evaluation. *Med Decis Making* 1996;16:309-14.
- Winkens RA, Pop P, Grol RP, Bugter Maessen AM, Kester AD, Beusmans GH, Knottnerus JA. Effects of routine individual feedback over nine years on general practitioners' requests for tests. *BMJ* 1996;312:490.
- Wirtschafter DD, Sumners J, Jackson JR, Brooks CM, Turner M. Continuing medical education using clinical algorithms. A controlled trial assessment of effect of neonatal care. *Am J Diseased Children* 1986;140:791-7.
- Wones RG. Failure of low-cost audits with feedback to reduce laboratory test utilization. *Med Care* 1987;25:78-82.
- Zaat JO, van Eijk JT, Bonte HA. Mag het ook een testje minder? De invloed van een beperking van het aanvraagformulier voor laboratoriumonderzoek. *Huisarts Wet* 1991;34:72-7.
- Zaat JO, van Eijk JT, Bonte HA. Laboratory test form design influences test ordering by general practitioners in the Netherlands. *Med Care* 1992;30:189-98

CHAPTER IV

Effect of a practice-based strategy on test ordering performance of primary care physicians. A randomized trial.

Wim HJM Verstappen
Trudy van der Weijden
Jildou Sijbrandij
Ivo Smeele
Jan Hermsen
Jeremy Grimshaw
Richard PTM Grol

Published in JAMA: Journal Of The American Medical Association 2003:289: 2407-12

A revised version was also published as:

Verstappen WH, van der Weijden T, Sijbrandij J, Smeele I, Hermsen J, Grimshaw J, Grol RPTM Diagnostisch toetsoverleg (DTO) vermindert overbodig gebruik van aanvullende diagnostiek door huisartsen. Huisarts Wet 2004; 47: 27-32

Abstract

Context

Numbers of diagnostic tests ordered by primary care physicians are growing and many of these tests seem to be unnecessary according to established, evidence-based guidelines. An innovative strategy that focused on clinical problems and associated tests was developed.

Objective

To determine the effects of a multifaceted strategy aimed at improving the performance of primary care physicians' test ordering.

Design

Multicenter, randomized controlled trial with a balanced, incomplete block design and randomization at group level. Thirteen groups of primary care physicians underwent the strategy for 3 clinical problems (arm A; cardiovascular topics, upper and lower abdominal complaints), while 13 other groups underwent the strategy for 3 other clinical problems (arm B; chronic obstructive pulmonary disease and asthma, general complaints, degenerative joint complaints). Each arm acted as a control for the other.

Setting

Primary care physician groups in 5 regions in the Netherlands with diagnostic centers recruited from May to September 1998.

Study Participants

Twenty-six primary care physician groups, including 174 primary care physicians.

Intervention

During the 6 months of intervention, physicians discussed 3 consecutive, personal feedback reports in 3 small group meetings, related them to 3 evidence-based clinical guidelines, and made plans for change.

Main Outcome Measure

According to existing national, evidence-based guidelines, a decrease in the total numbers of tests ordered per clinical problem, and of some defined inappropriate tests, is considered a quality improvement.

Results

For clinical problems allocated to arm A, the mean total number of requested tests per 6 months per physician was reduced from baseline to follow-up by 12% among physicians in the arm A intervention, but was unchanged in the arm B control, with a mean reduction of 67 more tests per physician per 6 months in arm A than in arm B ($P = .01$). For clinical problems allocated to arm B, the mean total number of requested tests per 6 months per physician was reduced from baseline to follow-up by 8% among physicians in the arm B intervention, and by 3% in the arm A control, with a mean reduction of 28 more tests per physician per 6 months in arm B than in arm A ($P = .22$). Physicians in arm A had a significant reduction in mean total number of inappropriate tests ordered for problems allocated to arm A, whereas

the reduction in inappropriate test ordered physicians in arm B for problems allocated to arm B was not statistically significant.

Conclusion

In this study, a practice-based, multifaceted strategy using guidelines, feedback, and social interaction resulted in modest improvements in test ordering by primary care physicians.

Introduction

In many countries, the number of diagnostic tests ordered by primary care physicians is growing, while according to established evidence-based guidelines, many of these tests are seen as unnecessary.[1-3] Possible explanations are test ordering routines that are difficult to change, a more defensive attitude among primary care physicians out of fear of medical errors, or a lack of knowledge about the appropriate use of tests.[4-7] Moreover, patients more actively ask for tests and often attach greater value to test results than is justified by the facts.[8-9] Unfortunately, little is yet known about the negative effects of performing such tests, in terms of, for example, unnecessary exposure to radiation or false-positive results, that may induce fear and anxiety in patients or may result in a cascade of unnecessary further testing.

Given these problems it is challenging to learn how to change test ordering performance effectively and bring it into line with existing evidence or guidelines on optimal testing. Many such attempts have been made with mixed results, showing that successful strategies require a well-balanced combination of interventions.[10-12] We have developed a multifaceted strategy combining personal feedback and guideline dissemination with quality meetings in small groups of primary care physicians. Social interactions were used as an important motivator for change, as physicians learned how colleagues were handling test ordering problems and as they obtained information about the consequences of medical decision making in daily practice.[13-14]

The aim of this strategy was to achieve sustained improvements in test ordering, for example, working in line with the national, evidence-based guidelines. The present article describes the changes in test ordering performance resulting from this innovative strategy in a large population of primary care physicians.

Methods

Setting and Population

Our study was conducted in 5 regions in the Netherlands, each of which made use of the services of a diagnostic center. A diagnostic center is an institute, usually associated with a hospital, where primary care physicians can order tests without referring patients to the hospital. Thirty-seven local groups of primary care physicians linked to 1 of these 5 diagnostic centers were eligible for the study. These groups are a common feature of Dutch general practice, involving teams of primary care physicians collaborating in a specific region. These teams share patient care outside office hours and many of them also engage in continuing medical education. From May until September 1998 the coordinators of the 5 diagnostic centers recruited local groups in their regions to participate.

Intervention

The strategy consisted of the following elements: personalized graphical feedback, including a comparison of each physician's own data with those of colleagues; dissemination of national, evidence-based guidelines; and regular meetings on quality improvement in small groups. The strategy focused on specific clinical problems and the diagnostic tests used for these problems (**Table 1**). These tests covered about 90% of all tests a primary care physician can order in a diagnostic center. For the tests used in the trial, national guidelines for optimal test ordering had to be available.

During the first 6 months of 1999, each of the recruited physicians received by mail 3 consecutive feedback reports on 3 different clinical problems, together with concise information on the 3 evidence-based clinical guidelines for these problems, developed by the Dutch College of Primary Care Physicians.

Each postal contact was followed by a 90-minute standardized small group quality improvement meeting about 2 weeks later, supervised by the medical coordinator of the diagnostic center. At the 3 meetings, physicians were asked to discuss and compare their feedback reports with colleagues and to relate them to the national guidelines. They also discussed Bayesian decision rules to help them understand the probability of false-positive results in low-prevalence disorders. Another important topic of debate was how to deal with the frequent requests by patients to have inappropriate tests performed. This discussion of the guidelines was followed by a thorough discussion of the difficulties of achieving changes at the individual primary care physician level, the practice level, or at the patient level. The next step was to try to implement the guidelines in their own practice, and at the end of each session, plans were drawn up for change, both at individual and group level. Subsequent meetings were used to evaluate whether targets had been met.

Design and Measurements

The effect of the intervention was evaluated in a multicenter, randomized controlled trial that was conducted in the first 6 months of 1999 with a balanced, incomplete block design, consisting of 2 arms, with the local group of primary care physicians as the unit of randomization (**Figure 1**). One group of local groups (arm A) underwent the strategy with respect to tests associated with the 3 clinical problems allocated to arm A (**Table 1**), while the other group of local groups (arm B) underwent the strategy with respect to tests associated with the 3 problems allocated to arm B (**Table 1**). The groups in arm A acted as blind controls for the groups undergoing the arm B intervention, and vice versa. This rigorous design was used to balance the influence of nonspecific effects on the test ordering performance between the 2 arms and to neutralize the Hawthorne effect, that is, the effect that physicians might change their test ordering because they were aware of taking part in a trial.[15-16] After stratification for region and group size, randomization was performed centrally with Duploran, a random numbers program. The physicians gave informed consent for the retrieval of anonymous data on the numbers and results of all tests ordered. To avoid seasonal influences, the numbers of tests for effect evaluation were assessed during the last 6 months of 1998 (the baseline period) and the last 6 months of 1999 (the follow-up period).

TABLE 1 CLINICAL PROBLEMS AND DIAGNOSTIC TESTS USED IN THE TRIAL.

	CLINICAL PROBLEMS / TESTS ARM A		CLINICAL PROBLEMS / TESTS ARM B
A1	**Cardiovascular diseases** Cholesterol, subfractions, potassium, sodium, creatinine, ECG (exercise), BUN*	B1	**COPD/Asthma** Allergic screening test, chest radiography, immonoglobulin E*
A2	**Upper abdominal complaints** SGPT, γ-glutamyltransferase, ultrasound scans of hepatobilliairy tract, SGOT*, LDH*, amylase*, bilirubin*, alkaline phosphatase*	B2	**General malaise / Vague complaints** ESR, Hb with or without indices, Ht, TSH, monospot, leucocyte count*
A3	**Lower abdominal complaints** Prostate-specific antigen, CRP, ultrasound of the kidney, IVP, double contrast barium enema, sigmoidoscopy	B3	**Degenerative joint complaints** ESR, uric acid, rheumatoid factors, X-rays of lumbar spine*, cervical spine*, shoulder*, knee*, hip*

* Tests that are inappropriate according to the national evidence-based guidelines

Intervention Effect Measures

Characteristics of primary care physicians and local groups were collected by means of a written questionnaire. Two effect measures were used to evaluate intervention effects:

1. A decrease in the total numbers of requested tests per 6 months per physician: since most of the recommendations in the national, evidence-based guidelines advise ordering fewer tests, a decrease in the total numbers of tests ordered was regarded as an improvement in patient care. Separate analyses were performed for the 6 different clinical problems.
2. A decrease in the numbers of inappropriate tests as defined in the guidelines (**Table 1 and Box I**): these tests were regarded as inappropriate for the associated clinical problems for various reasons, for example, because the results of these tests seldomly have an influence on the treatment, because the high likelihood of false-positive results can occur, because better alternatives are available, or because adverse effects to some tests can occur (eg, radiology tests).

Statistical Analysis

Differences in individual characteristics of the primary care physician were tested for significance with Pearson χ^2 test. In the evaluation of intervention effects, the unit had to be the local group of primary

care physicians because this unit was also the unit of randomization. To account for clustering within local groups, a 3-level model was used with the local group as level 3, individual physicians as level 2, and numbers of tests as level 1. The analysis was carried out with SAS PROC MIXED, release 8.2 (SAS Institute, Cary, NC). Power calculations based on the baseline data showed that each arm needed approximately 85 physicians to detect a 10% difference in mean total numbers of tests with 80% power, and a risk of type 1 error of .05. All effects were analyzed with analysis of covariance using the numbers of tests during the follow-up period as the dependent variable and the numbers of tests at baseline and the region, which appeared to be an important determinant, as independent variables.

Results

One hundred seventy-four primary care physicians, belonging to 26 local groups, expressed their willingness to participate on first request, so no further recruitment was necessary. After randomization, both arms included 13 local groups (**Figure 1**). No differences were found among the characteristics of our individual study primary care physicians (**Table 2**). Likewise, no differences were found in the characteristics of the local primary care physician groups (data not shown). The mean size of the local groups and experience with continuing medical education in small groups of colleagues did not differ between the 2 arms, nor was there any statistically significant difference

TABLE 2 STUDY POPULATION CHARACTERISTICS AT PRIMARY CARE PHYSICIAN LEVEL

	ARM A	ARM B
No. of physicians	85	89
Age, mean(SD), y	46.2 (6.6)	45.8 (5.4)
Female, No. (%)	14 (16)	15 (17)
Mean no.(SD) of patients per physician*	2587 (641)	2637 (519)
Patients > 65y, % mean (SD)	15 (6.8)	13 (7.1)
Working time factor**, %, (SD)	91 (15)	91 (16)
Physicians with a solo practice, No.(%)	43 (51)	48 (54)
Physicians who use computerized registration system, No. (%)	66 (78)	61 (69)

* Total practice population for whom the primary care physician is responsible. ** Part-time factor is the working time. A full-time factor is 100%, each half of the day is 10%, so the part-time factor of 80% is a physician who works 4 days.

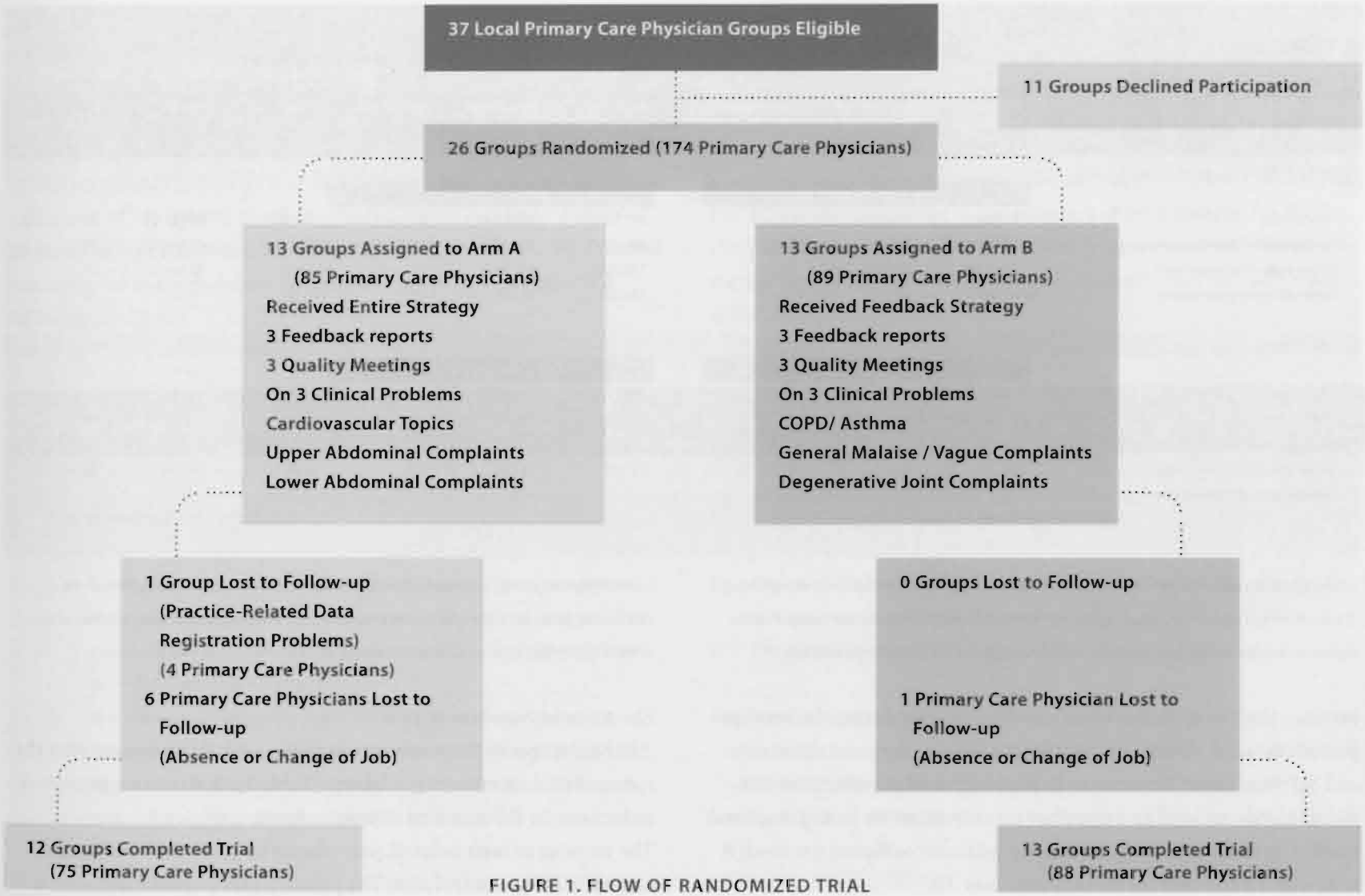


FIGURE 1. FLOW OF RANDOMIZED TRIAL

TABLE 3 EFFECTS OF THE STRATEGY BY ANALYSIS OF COVARIANCE ADJUSTED FOR NUMBERS OF DIAGNOSTIC TESTS AT BASELINE AND FOR THE REGION ON THE MEAN (SD) NUMBERS OF TESTS, PER PRIMARY CARE PHYSICIAN PER 6 MONTHS

CLINICAL PROBLEM	BASELINE MEAN (SD)	FOLLOW-UP MEAN (SD)	% CHANGE	BASELINE MEAN (SD)	FOLLOW-UP MEAN (SD)	% CHANGE	β (SE)*	95% CI	P VALUE
ARM A TESTS									
	ARM A (INTERVENTION)			ARM B (CONTROL)					
TOTAL TESTS	478 (309)	422 (234)	-12	507 (293)	503 (281)	0	-67 (19)	-104 to -30	.01
Cardiovascular/hypertension	293 (189)	276 (157)	-6	290 (182)	302 (184)	+4	-35 (13)	-61 to -10	.01
Upper abdominal complaints	165 (125)	128 (82)	-22	192 (128)	174 (114)	-9	-28 (9)	-45 to -10	.01
Lower abdominal complaints	20 (20)	18 (19)	-10	25 (25)	27 (29)	+8	-5 (2)	-9 to -1	.02
ARM B TESTS									
	ARM A (CONTROL)			ARM B (INTERVENTION)					
TOTAL TESTS	640 (394)	624 (357)	-3	724 (386)	664 (356)	-8	-28 (23)	-74 to 14	.22
COPD/asthma	39 (31)	31 (25)	-20	53 (27)	38 (19)	-28	-1 (2)	-5 to 3	.58
General complaints	548 (340)	544 (310)	0	599 (340)	568 (321)	-5	-19 (21)	-61 to 22	.36
Lower abdominal complaints	54 (38)	49 (36)	-9	72 (43)	58 (37)	-19	-3 (4)	-10 to 4	.34
Degenerative joint compliants									

Abbreviations: CI, confidence interval; COPD, chronic obstructive pulmonary disease; SE, standard error.

*β is the intervention effect (analysis of covariance) from which the follow-up numbers of tests are the dependent variable and the baseline numbers and the region are the independent variables. β reflects the total change between baseline and follow-up in mean (SD) numbers of tests in the intervention group minus the total change between baseline and follow-up in mean numbers of tests in the control group, adjusted for baseline and region.

between the 2 arms in the mean numbers of tests during the baseline period (data not shown). In multilevel analyses, the point estimation and SD were about the same as in the analysis of covariance at individual physician level and therefore no correction for local groups was needed, even though the intraclass correlation coefficient for block A tests was .12 and that for block B tests was .10.

Decreases in Numbers of Tests

All the changes in the intervention group were in agreement with the national evidence-based guidelines (**Table 3**), that is, the represented reductions in the numbers of tests ordered.

The number of tests ordered were always larger in the intervention arm than in the control arm. The primary care physicians in arm A

decreased the total mean numbers of tests relating to problems allocated to arm A by 12% between baseline and follow-up, while no change in the numbers of these tests occurred for primary care physicians in arm B (blind control arm). The decrease for physicians in arm A was 67 tests more per physician compared with the decrease for the physicians in arm B (P = .01). The physicians in arm B achieved a decrease of 8% in total number of tests ordered for the problems allocated to arm B between baseline and follow-up, while a 3% decrease was achieved in the numbers of these tests by physicians in arm A (blind control arm).

These results correspond with an additional decrease in the total numbers of tests for problems allocated to arm B of 28 compared with the physicians of arm A (P = .22). The results per clinical problem also are shown in **Table 3**. The mean change in numbers of tests ordered for the 3 clinical problems allocated to arm A was statistically significant (cardiovascular, P = .01; upper abdominal, P = .01; lower abdominal, P = .02), while the change in the numbers of tests ordered for the 3 clinical problems allocated to arm B was in agreement with the recommendations in the national guidelines, although each failed to reach statistical significance.

Inappropriate tests as defined in evidence-based guidelines BOX 1

Upper abdominal complaints

1. There is no reason to order liver function tests for vague upper abdominal complaints without jaundice. The risk of false-positive results is too large because of the low prevalence of patients with liver diseases in general practice
2. If screening is necessary, order serum glutamic-pyruvate transaminase and γ-glutamyltransferase in patients without jaundice
3. Order total bilirubin, serum glutamic-pyruvate transaminase, and γ-glutamyltransferase in patients with jaundice

General malaise, fatigue, and vague complaints

1. Order hemoglobin and erythrocyte sedimentation rate in patients with general fatigue that has persisted for longer than 1 month
2. Do not order leukocyte counts in cases of general fatigue

Degenerative joint complaints

Do not order radiographs of the joints since the results of these tests have no influence on the treatment

Decreases in Numbers of Inappropriate Tests

The reduction in the total numbers of inappropriate tests is shown in **Table 4**. After the intervention, significantly fewer total inappropriate tests for the problems allocated to arm A were ordered by the primary care physicians in this arm (P = .01). The total numbers of inappropriate tests for the problems allocated to arm B ordered by the primary care physicians in arm B also tended to decrease, which was in agreement with the recommendations in the guidelines, but the reduction failed to reach statistical significance (P = .11). A significant reduction in the numbers of tests ordered, compared with the control group, was found for 4 of the tests for upper abdominal complaints: amylase, bilirubin, lactic dehydrogenase, and alkaline phosphatase.

TABLE 4 EFFECTS OF THE STRATEGY BY ANALYSIS OF COVARIANCE ADJUSTED FOR NUMBERS OF DIAGNOSTIC TESTS AT BASELINE AND FOR THE REGION ON THE MEAN (SD) NUMBERS OF INAPPROPRIATE TESTS, PER PRIMARY CARE PHYSICIAN PER 6 MONTHS

CLINICAL PROBLEM	BASELINE MEAN (SD)	FOLLOW-UP MEAN (SD)	BASELINE MEAN (SD)	FOLLOW-UP MEAN (SD)	β (SE)*	95% CI	P VALUE
ARM A TESTS							
	ARM A (INTERVENTION)		ARM B (CONTROL)				
TOTAL TESTS	63 (75)	45 (41)	66 (55)	63 (56)	-16 (4.8)	-27 to 07	.01
BUN	8.7 (19)	7.2 (15)	6.3 (7.2)	6.6 (8.3)	-1 (1.3)	-4 to 2	.37
SGOT	7.7 (11)	5.5 (7.7)	8.3 (13)	7.3 (14)	-2 (1.4)	-5 to 1	.13
LDH	13 (27)	8.8 (16)	12 (20)	11 (18)	-3 (1.5)	-6 to -1	.01
Amylase	5.3 (13)	3.6 (6.9)	3.4 (4.9)	4.5 (10)	-2 (1.1)	-4 to -0.1	.04
Alkaline phosphatase	11 (25)	7.0 (11)	9.3 (13)	9.0 (15)	-3 (1.5)	-6 to -0.3	.03
Bilirubin	20 (27)	15 (19)	31 (43)	27 (35)	-6 (2.6)	-11 to -0.3	.04
ARM B TESTS							
	ARM A (CONTROL)		ARM B (INTERVENTION)				
TOTAL TESTS	134 (81)	126 (74)	163 (89)	138 (74)	-8 (5.0)	-18 to 2	.11
Immunoglobulin E	3.6 (5.3)	2.8 (4.7)	3.0 (5.3)	1.5 (2.7)	-1 (0.42)	-1 to 1	.14
Leukocyte count	95 (63)	92 (57)	110 (69)	96 (58)	-6 (4.0)	-4 to 2	.11
Total imaging tests†	36 (26)	31(22)	50 (34)	41 (26)	-1 (2.7)	-4 to 6	.70

Abbreviations: CI, confidence interval; COPD, chronic obstructive pulmonary disease; BUN, blood urea nitrogen; LDH, lactic dehydrogenase; SE, standard error; SGOT, serum glutamic-oxaloacetic transaminase.

*See footnote in Table 3 for the intervention effect β.

†Total imaging tests include chest radiography, radiographs of the lumbar spine, cervical spine, shoulder, knee and hip.

Comment

A new strategy to influence test ordering performance was evaluated in a trial with a large group of primary care physicians in 5 diagnostic center regions in the Netherlands.

The relatively short intervention period resulted already in a substantial reduction in the total numbers of tests ordered and in the number of inappropriate tests ordered. Although the effects may seem not very large, it is important to realize primary care physicians in the Netherlands already order fewer tests than their colleagues in other countries.[1] This further reduction can be regarded as quality improvement in terms of test ordering because these changes were in agreement with the recommendations in national evidence-based guidelines.

There are some methodological considerations. We have no reason to believe that the large study population differs from the Dutch primary care physician population. Items relevant for the determinants of test ordering performance of primary care physicians were distributed equally over both arms.[17] However, maybe only motivated, well-functioning groups of physicians participated, and it is therefore questionable if the strategy will work for all groups. Secondly, our study only evaluated effects on volume of tests, because patient data were not available from the diagnostic centers. However, available empirical evidence shows that a general reduction in test ordering in primary care does not lead to more referrals or substitution of care.[18] Furthermore, despite that the guidelines state that a reduction in total test ordering equals quality improvement, this does not implicate that each separate test should always decrease. Finally, the duration of the study is too short to determine long-term effects on test ordering.

Our study underlines that multifaceted interventions are superior to single interventions.[19-20] Significant changes in numbers of tests were not found for all clinical problems included, so conclusions about the effectiveness of our strategy are not straightforward. Some clinical problems may require additional strategies, for example, electronic reminders may be necessary to achieve further improvement.[21] Nevertheless, our strategy would seem to be a powerful effective and tailor-made strategy, which fits in well with routine primary care physician practice in many western countries, is linked to the every day general practice routine, and gives primary care physicians the opportunity to discuss their test ordering performance with colleagues on the basis of actual performance data, making discussions less non-committal. Discussing feedback reports and guidelines provides physicians the opportunity to change their performance by learning from each other and by learning to implement new strategies. Thus, social influence by peer interaction can be an important motivator for change.[14] Our strategy could also be used for in-hospital teams or other groups of collaborating physicians, as well as for other topics, such as prescription or referral behavior.

Funding/Support:
This study was supported by the Dutch Health Care Insurance Council.

References of studies in the text

1. Leurquin P, Van Casteren V, De Maeseneer J. Use of blood tests in general practice: a collaborative study in eight European countries: Eurosentinel Study Group. *Br J Gen Pract.* 1995;45:21-25.
2. Ayanian JZ, Berwick DM. Do physicians have a bias toward action? a classic study revisited. *Med Decis Making.* 1991;11:154-158.
3. Kristiansen IS, Hjortdahl P. The general practitioner and laboratory utilization: why does it vary? *Fam Pract.* 1992;9:22-27.
4. Ferrier B, Woodward C, Cohen M, et al. Clinical practical guidelines: new-to-practice family physicians' attitudes. *Can Fam Physician.* 1996;42:463-468.
5. Zaat JO, van Eijk JT. General practitioners' uncertainty, risk preference, and use of laboratory tests. *Med Care.* 1992;30:846-854.
6. Wong ET. Improving laboratory testing: can we get physicians to focus on outcome? *Clin Chem.* 1995;41:1241-1247.
8. Hoffrage U, Lindsey S, Hertwig R, et al. Medicine: communicating statistical information. *Science.* 2000;290:2261-2262.
8. McDonald IG, Daly J, Jelinek VM, et al. Opening Pandora's box: the unpredictability of reassurance by a normal test result. *BMJ.* 1996;313:329-332.
9. Little P, Cantrell T, Roberts L, et al. Why do GPs perform investigations? the medical and social agendas in arranging back x-rays. *Fam Pract.* 1998;15:264-265.
10. Grimshaw JM, Russell IT. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. *Lancet.* 1993;342:1317-1322.
11. Oxman AD, Thomson MA, Davis DA, et al. No magic bullets: a systematic review of 102 trials of interventions to improve professional practice. *CMAJ.* 1995;153:1423-1431.
12. Solomon DH, Hashimoto H, Daltroy L, et al. Techniques to improve physicians' use of diagnostic tests: a new conceptual framework. *JAMA.* 1998;280:2020-2027.
13. Grol R. Peer review in primary care. *Qual Assur Health Care.* 1990;2:119-126.
14. Mittman B, Tonesk X, Jacobson P. Implementing clinical practice guidelines: social influence strategies and practitioner behavior change. *Qual Rev Bull.* 1992:413-422.
15. Brady WJ, Hissa DC, McConnell M, et al. Should physicians perform their own quality assurance audits? *J Gen Intern Med.* 1988;3:560-565.
16. Winkens RA, Pop P, Bugter Maessen AM, et al. Randomised controlled trial of routine individual feedback to improve rationality and reduce numbers of test requests. *Lancet.* 1995;345:498-502.
17. Bugter Maessen AM, Winkens RA, Grol RP, et al. Factors predicting differences among general practitioners in test ordering behaviour and in the response to feedback on test requests. *Fam Pract.* 1996;13:254-258.
18. Winkens RA, Grol RP, Beusmans GH, et al. Does a reduction in general practitioners' use of diagnostic tests lead to more hospital referrals? *Br J Gen Pract.* 1995;45:289-292.
19. Wensing M, Grol R. Single and combined strategies for implementing changes in primary care: a literature review. *Int J Qual Health Care.* 1994;6:115-132.
20. Wensing M, Van der Weijden T, Grol R. Implementing guidelines and innovations in general practice: which interventions are effective? *Br J Gen Pract.* 1998;48:991-997.
21. Eccles M, Steen N, Grimshaw J, et al. Effect of audit and feedback, and reminder messages on primary-care radiology referrals: a randomised trial. *Lancet.* 2001;357:1406-1409

CHAPTER V

Improving test ordering in primary care: the added value of a small group quality improvement strategy over classic feedback only. A multicenter randomized trial.

Wim HJM Verstappen
Trudy van der Weijden
Willy I Dubois(†)
Ivo Smeele
Jan Hermsen
Frans ES Tan
Richard PTM Grol

Published in Annals of Family Medicine 2004;2 (In press)

Abstract

Background

Numbers of tests ordered by primary care physicians (PCP) are growing and many of these tests seem to be unnecessary according to established, evidence-based guidelines.

Objective

Evaluation of the added value of small peer group quality improvement meetings compared with simple feedback as a strategy to improve test ordering behavior.

Design

Cluster randomized trial with randomization at local primary care physician group level.

Setting

194 PCPs organized in 27 local PCP groups in 5 regions (5 diagnostic centers).

Intervention

An innovative, multifaceted strategy, combining written comparative feedback, group education on national guidelines and social influence by peers in quality improvement sessions in small groups.

Measurements

The mean number of tests per PCP per six months at baseline and the PCPs' region were used as independent variables, and the mean number of tests per PCP per six months as the dependent variable.

Results

The new strategy was executed in 13 PCP groups, while 14 groups received feedback only. In the intervention arm the decrease of the mean total number of tests was far more substantial (on average 51 tests less per PCP per half year) compared with the feedback arm ($p=0.0049$). Five 'inappropriate' tests for the clinical problem 'upper abdominal complaints' decreased in the intervention arm with 13 tests more per PCP per 6 months than in de feedback arm ($p=0.0015$). Inter-doctor variation decreased more in the intervention arm.

Conclusions

Compared to only disseminating comparative feedback reports to PCPs, the new strategy, involving peer interaction and social influence, improved the PCPs' test ordering behavior. In order to be effective, feedback needs to be integrated in an interactive, educational environment.

Key words:

Quality assurance, health care; test ordering behavior; feedback; small group quality improvement.

Acknowledgements

The authors gratefully acknowledge the financial contribution to the study provided by the Dutch College for Health Insurances.

Introduction

Numbers of tests ordered by primary care physicians (PCPs) are growing in many countries, and inter-doctor variation is shown to be large, while according to established guidelines many of these tests can be seen as unnecessary.[1-3] It is as yet unclear, however, what would be the best method to influence PCPs' test ordering behavior. Several studies evaluating different types of interventions to change this behavior have, so far, shown heterogeneous results. One of these widely investigated strategies with mixed results is feedback.[4-7] Many authorities in western countries, such as health insurers, regularly disseminate feedback reports about test ordering, prescription or referral rates to physicians or practices, often without substantial impact.[8,9]

The literature shows that multifaceted strategies in general are superior to single methods when it comes to influencing behavior.[10-12] Success rates of specific strategies seem to be strongly influenced by the extent to which they fit within the local and organizational context and the physicians' everyday work routine.[13,14] Favorable experiences have been gained particularly with small group education and interactive quality improvement sessions for primary care physicians.[15,16]

We therefore decided to develop a multifaceted strategy, combining transparent, individual graphical feedback on test ordering routines, education on clinical guidelines for test ordering and small group quality improvement meetings among PCPs, in which test ordering behavior and changes in routines are discussed, using social influence and peer influence as important motivators of change. Social influence from respected colleagues or opinion-leaders seem to have more effect on changing practice routines than traditional medical education activities, focusing on changing professional cognitions or attitudes.[17-21] Therefore, our strategy seemed promising, since it is closely linked to the everyday setting of many PCPs, who tend to work more or less in isolation and have limited contact with peers about subjects like test ordering behavior.

In a multicenter randomized trial with a block design this strategy actually had a favorable effect on the test ordering behavior of PCPs.[22] Since classic feedback is an increasingly routine quality improvement strategy, we were interested to assess the added value of this innovative, multifaceted strategy compared with standardized feedback only, one of the elements of the strategy.

Methods

Overall design and population

A multicenter RCT was conducted during the first six months of 1999 in five regions with a diagnostic center. A diagnostic center is an institute, usually associated with a hospital, where PCPs can order laboratory, imaging and function tests. All five diagnostic centers used nationally developed indication-oriented forms for laboratory orders. In the 5 regions 37 local PCP groups with 294 PCPs were eligible for participation, since they made use of one of these five diagnostic centers. Local PCP groups are an existing part of the infrastructure of Dutch PCPs collaborating in a specific region, and sharing patient care outside office hours. Continuous medical education, for example by means of quality meetings about prescribing, is an important activity in most groups. One of the tasks of the medical coordinators of diagnostic centers is to give feedback to PCPs on their test ordering behavior, and they are

considered as opinion-leaders concerning test ordering. From May 1998 until September 1998 the coordinators of the five diagnostic centers recruited local PCP groups in their regions to participate in the trial.

Intervention

The new strategy consisted of the following elements: personalized graphical feedback, including a comparison of each PCP's own data with those of colleagues, dissemination of and education on national, evidence-based guidelines, and continuous quality improvement meetings in small groups. The improvement strategy concentrated on three specific clinical subjects (cardiovascular topics, upper abdominal complaints and lower abdominal complaints) and the tests used for these clinical problems, since it was felt that PCPs would prefer to discuss specific clinical topics rather than specific tests (**Table 1**).

During the first half year of 1999 each PCP received three different feedback reports (**Figure 1**) on these three clinical problems by mail, together with concise information on the evidence-based clinical guidelines for these specific clinical subjects, developed by the Dutch College of Primary Care Physicians. Each postal contact was followed by a 90-minute standardized small group quality improvement meetings about two weeks later, at which one of the clinical problems was discussed, based on the feedback reports and the guidelines (**Figure 2**). In these meetings social influence was an important vehicle to reach improvement on test ordering, and consisted of the following major components. The first was mutual personal feedback by peers, who worked in pairs at the start of the meeting. The second component was an interactive group education in which national guidelines were related to the individual PCPs' actual test ordering behavior, and to reach a kind of group consensus on the optimal test ordering

TABLE 1 CLINICAL PROBLEMS AND ASSOCIATED TESTS USED IN THE TRIAL

CLINICAL PROBLEMS / TESTS	
Cardiovascular topics Cholesterol, subfractions, potassium, sodium, creatinine, BUN, ECG (exercise)	**Upper abdominal complaints** SGPT, SGOT*, LDH*, amylase*, γ-glutamyltransferase, bilirubin*, alkaline phosphatase*, ultrasound scans of hepatobiliary tract
Lower abdominal complaints Prostate-specific antigen, CRP, ultrasound of the kidney, IVP, double contrast barium enema, sigmoidoscopy	

*Tests that are inappropriate according to national evidence-based guidelines on upper abdominal complaints. (see Box)

FIGURE 1. AN EXAMPLE OF A FEEDBACK REPORT

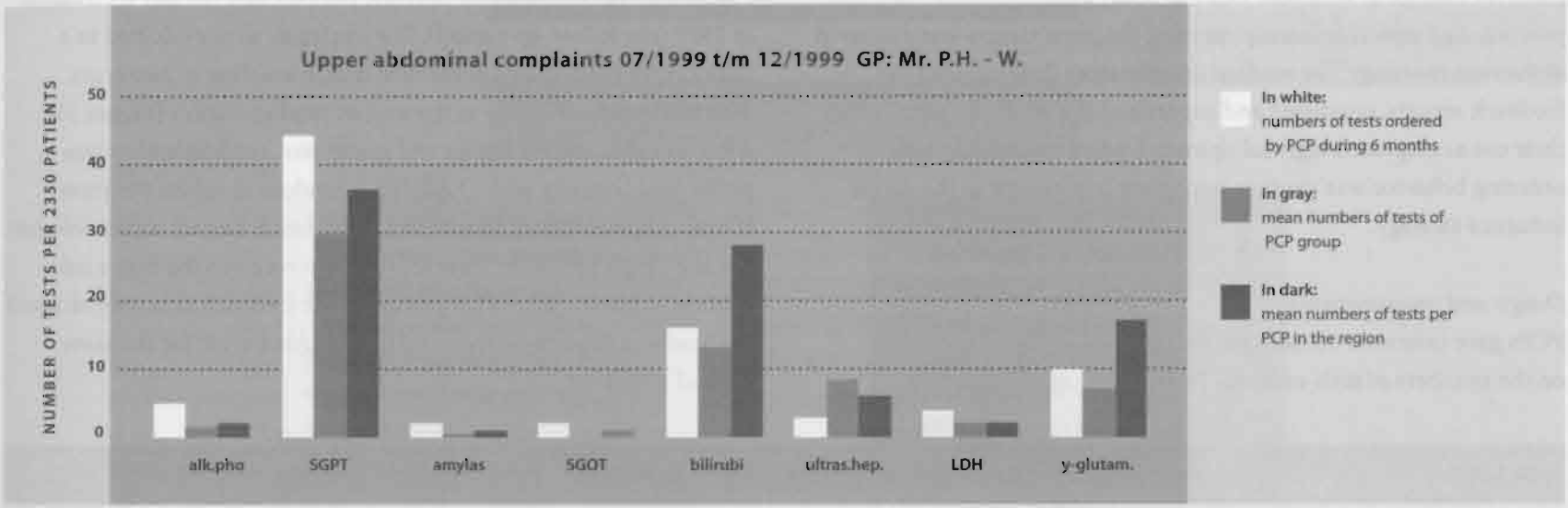
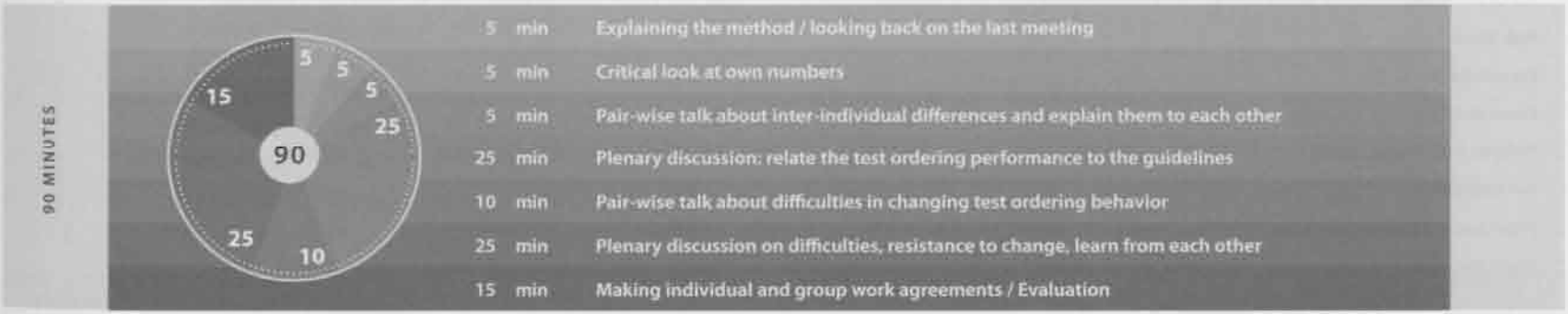


FIGURE 2. STRUCTURE OF THE 90-MINUTES SMALL GROUP QUALITY MEETNG



behavior. The third was the development of individual and group plans for change to stimulate PCPs to really put their plans into daily practice. As a critical follow-up, meeting the plans' targets was discussed at the next meeting. The medical coordinators disseminated the feedback reports, organized and supervised the quality meetings, and their use as respected regional opinion leaders concerning test ordering behavior was another important component in the social influence strategy.

Design and measurements

PCPs gave informed consent for the retrieval of anonymous data on the numbers of tests ordered. To avoid seasonal influences, the numbers of tests for effect evaluation were assessed over the last six months of 1998 (the baseline period), and the last six months of 1999 (the follow-up period). The strategies were evaluated in a multicenter randomized controlled trial, consisting of two arms, with the local PCP group as the unit of randomization (**Figure 3**). After stratification for region and group size, randomization was performed centrally with Duploran, a random numbers program. (Dept. of Epidemiology, Maastricht University, F. Kessels, methodologist). Local PCP groups of the intervention arm received the entire intervention, while the local PCP groups of the feedback arm only received the feedback reports on their test ordering behavior for the same clinical problems.

TABLE 2 STUDY POPULATION CHARACTERISTICS AT INDIVIDUAL PRIMARY CARE PHYSICIAN LEVEL

	INTERVENTION ARM	FEEDBACK ARM
No. of physicians	85	109
Age, mean(SD), y	46.2 (6.6)	46.2 (6.6)
Female, No. (%)	14 (16)	11 (10)
Mean no.(SD) of patients per physician*	2587 (641)	2444 (416)
Patients > 65y, % mean (SD)	15 (6.8)	15 (6.5)
Working time factor**, %, (SD)	91 (15)	92 (12)
Physicians with a solo practice, No.(%)	43 (51)	44 (40)
Physicians who use computerized registration system, No. (%)	66 (78)	75 (69)

* Total practice population for whom the primary care physician is responsible.

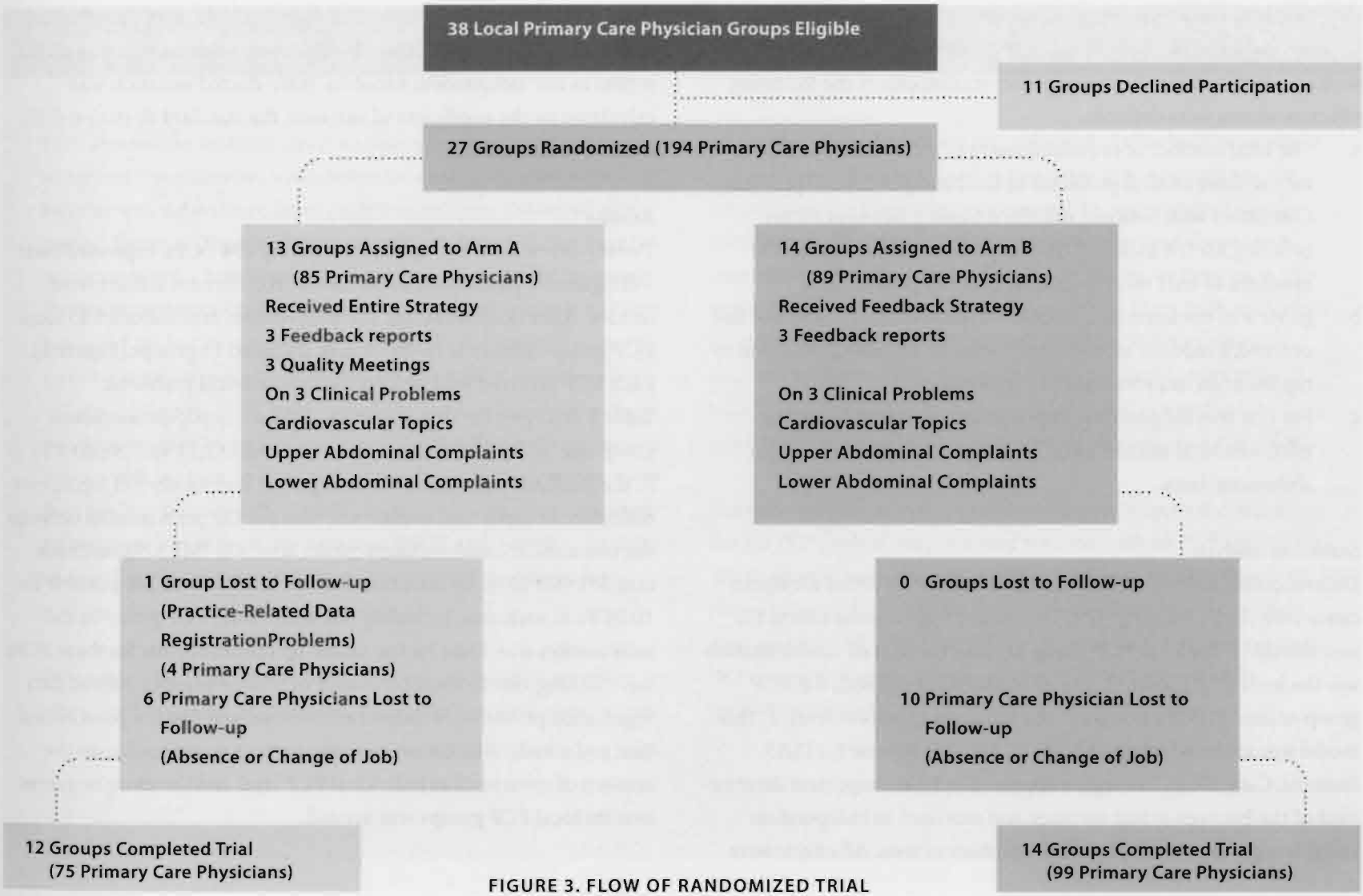


FIGURE 3. FLOW OF RANDOMIZED TRIAL

Effect measures and measuring instruments.
Characteristics of PCPs and local PCP groups were collected by a written questionnaire. To evaluate intervention effects the following effect measures were defined:

a. The total number of requested tests per six months per PCP for the three clinical problems in total and per clinical problem. Consistent with national, evidence-based guidelines for test ordering for the included clinical problems a decrease in the numbers of tests was considered as better patient care.
b. In view of the large inter-doctor variation in the numbers of test ordered, a reduced inter-doctor variation was also considered to represent an improvement in performance.
c. For one specific problem, upper abdominal complaints, the effects on total numbers and on defined inappropriate upper abdominal tests.

Statistical analysis
Differences on individual PCP characteristics were tested for significance with the Pearson's χ^2-test. To evaluate intervention effects the unit should be the local PCP group because the unit of randomization was the local PCP group. A 3-level model was used with the PCP group as level 3, PCPs as level 2, and numbers of tests as level 1. This model was analyzed using SAS PROC MIXED Release 8.2 (SAS Institute, Cary, NC). The region appeared to be an important determinant of the between group variance and was used as independent variable together with the baseline numbers of tests. All effects were analyzed with analysis of covariance with the follow-up numbers of tests as dependent variable and the baseline numbers of tests and the region as the independent variables. Inter-doctor variation was calculated by the coefficient of variance, the standard deviation (SD) divided by the mean.

Results
Twenty-seven local PCP groups, including 194 PCPs, expressed their willingness to participate, so no further recruitment actions were needed. After randomization, the intervention arm included 13 local PCP groups, while the feedback arm included 14 groups (**Figure 3**). Each PCP received feedback on the three clinical problems.
Table 2 describes the characteristics of the study population. Mean group size in the intervention arm was 6.9 (SD 2.1), vs. 7.8 (SD 4.2) in the feedback arm. There was a large, but statistically not significant difference in mean total numbers of tests per GP per 6 months between the two arms at baseline; intervention arm: 478 (SD 309), feedback arm 541 (SD 337). An intention-to-treat analysis was not possible for 10 PCPs in each arm, including one entire local PCP group in the intervention arm. Data for the follow-up measurements for these PCPs were lacking, due to absence, change of jobs or practice-related data registration problems. Multilevel analyses showed that the point estimation and standard deviation were the same at group level as in the analysis of covariance at individual PCP level and therefore no correction for local PCP groups was needed.

Upper Abdominal Complaints **BOX 1**

PCPs received the feedback report on tests ordered in case of upper abdominal complaints two weeks before the small group quality meeting, together with the evidence-based guidelines on upper abdominal complaints (**Figure 1**). These guidelines recommend, first of all, that there is no reason to order liver function tests for non-specific upper abdominal complaints without jaundice. The risk of false-positive results is too large, because of the low prevalence of patients with liver diseases in primary care (4-5 per 1000 patients). If PCPs think screening is necessary, they are advised to order SGPT and γ-glutamyltransferase in patients without jaundice, and to order total bilirubin, SGPT and γ-glutamyltransferase in patients with jaundice. In short, there is never an indication to order more than two liver function tests in patients with upper abdominal complaints without jaundice, so the following 5 tests: SGOT, LDH, amylase, bilirubin and alkaline phosphatase, are seen as 'inappropriate' for patients with non-specific upper abdominal complaints.

At the meetings, PCPs discussed their reports, compared them with each other's results and with the guidelines, and also discussed Bayesian decision rules to help them understand the probability of false-positive results in low-prevalence disorders. Another important topic of debate was the frequent requests by patients with non-specific upper abdominal complaints to have blood tests. It took quite some effort and discussion to convince the PCPs they had to change their routine for these cases. The next step was to try and implement the guidelines. Many PCPs made plans for changes on this item, such as 'I will order less liver function tests, because I understand that these tests do not add useful information to what I know'. Some local PCP groups stated that they would use the same information brochure about non-specific upper abdominal complaints.

The intervention arm PCPs ordered on average 24 'upper abdominal test' less per PCP per half year, compared with the feedback PCP ($p=0.0031$). The number of 'inappropriate' tests for this clinical problem decreased from 55 (SD 60) to 39 (SD 32), while in the feedback arm the number decreased from 60 (SD 63) to 56 (SD 54), meaning that the intervention PCPs ordered 13 inappropriate tests less than the feedback PCPs ($p=0.0015$) (**Table 3**).

Table 3 shows results of these analyses at individual PCP level for all tests and per clinical problem. The total number of tests ordered decreased in both arms. For the intervention group PCPs the decrease was 51 tests more per PCP per half year than for the feedback PCPs (p=0.0049). The differences in changes were significant, except for cardiovascular topics that decreased with marginal significance. The Box I describes the intervention and its effects in more detail for the clinical problem "upper abdominal complaints": the differences for the defined inappropriate tests were also significant, meaning that the intervention PCPs ordered 13 inappropriate tests less than the feedback PCPs per PCP per half year (p=0.0015). **Table 3** also shows that the coefficient of variance decreased more in the intervention arm, meaning that the variation in test-ordering between intervention PCPs decreased more than in the feedback arm. **Figure 4** depicts the results for all tests at aggregated local PCP group level in graphical format, and shows that effects in the intervention arm were more straightforward.

TABLE 3 EFFECTS OF THE STRATEGY BY ANALYSIS OF COVARIANCE ADJUSTED FOR NUMBERS OF DIAGNOSTIC TESTS AT BASELINE AND FOR THE REGION ON THE MEAN (SD) NUMBERS OF TESTS, PER PRIMARY CARE PHYSICIAN PER 6 MONTHS

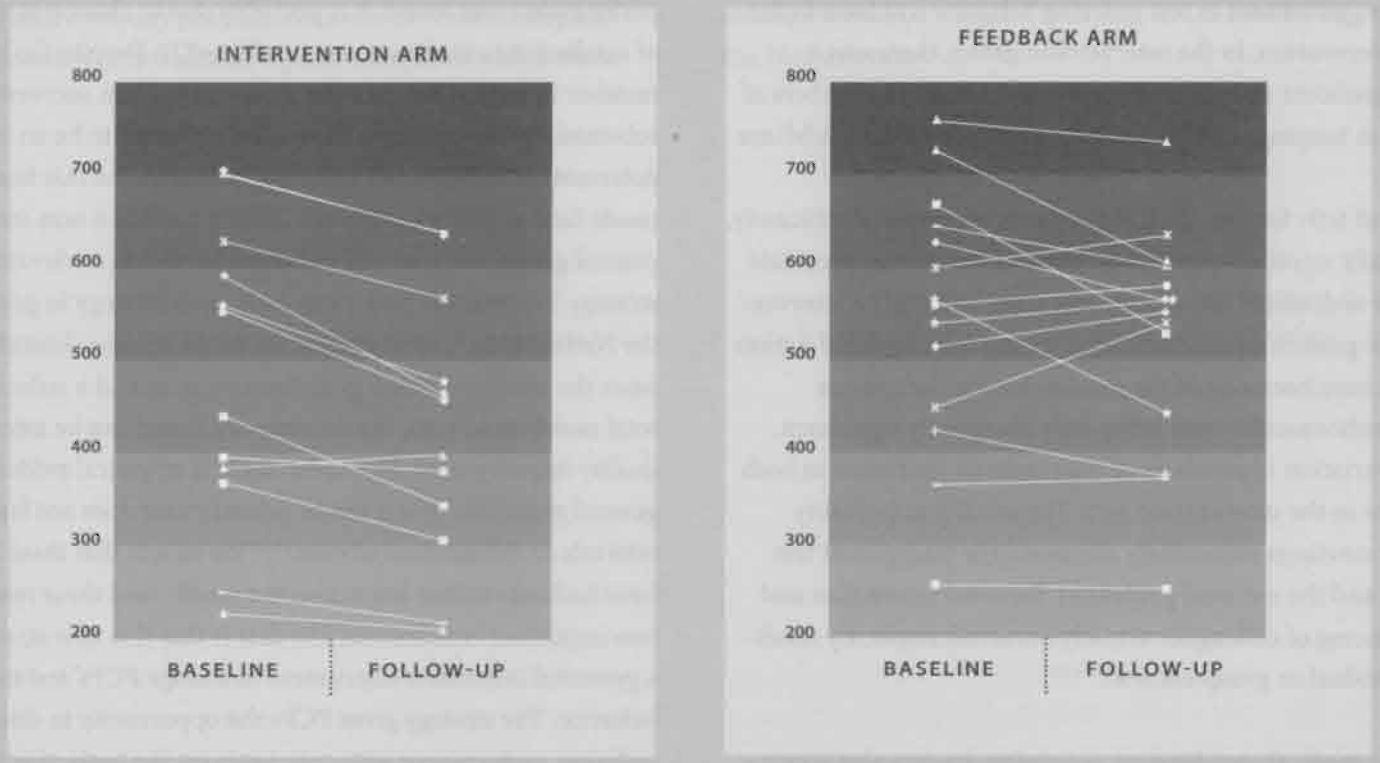
CLINICAL PROBLEM	INTERVENTION ARM				FEEDBACK ARM				B	SE. B	95% CI.	P
	BASELINE	CV	FOLLOW-UP	CV	BASELINE	CV	FOLLOW-UP	CV				
Total number of tests	478 (309)	0.65	422 (235)	0.56	541 (337)	0.62	535 (309)	0.58	-51	17.94	-87; -16	.0049
Cardiovascular topics	293 (189)	0.65	276 (157)	0.57	322 (214)	0.66	333 (205)	0.62	-25	13.08	-51; 1	.0561
Lower abdominal complaints	20 (20)	1.00	18 (19)	1.06	30 (40)	1.43	30 (27)	0.90	-6	2.18	-10; -2	.0076
Upper abdominal complaints	165 (125)	0.76	128 (82)	0.64	188 (143)	0.76	171(117)	0.68	-24	7.98	-40; -8	.0031
Inappropriate upper abdominal tests	55 (60)	1.09	39 (32)	0.82	60 (63)	1.05	56 (54)	0.96	-13	4.1	-22;-5.2	.0015

Abbreviations: SD, standard deviation; SE, standard error; CV, Coefficient of variance.

β= intervention effect = the total change between baseline and follow-up of mean numbers of tests in the intervention group – total change of numbers between baseline and follow-up of mean numbers of tests in the feedback group

CV = SD / mean

FIGURE. 4 BASELINE AND FOLLOW-UP MEASUREMENTS IN MEAN TOTAL NUMBERS OF TESTS PER 6 MONTHS AT AGGREGATED LOCAL PCP GROUP LEVEL FOR THE 13 INTERVENTION AND THE 14 FEEDBACK LOCAL PCP GROUPS.



Discussion

A new interactive quality improvement strategy was evaluated and compared with classic feedback alone among 27 local PCP groups, including 194 PCPs, in 5 regions. The first success was the easy recruitment, with PCP groups anxious to participate in the trial. A considerable improvement in test ordering behavior was found after one year of intervention. In the intervention group, there was a statistically significant and clinically relevant decrease in numbers of tests ordered, in keeping with the national evidence-based guidelines.

The numbers of tests for two clinical problems improved significantly, and a statistically significant reduction in numbers of 'inappropriate' tests for upper abdominal complaints was seen. During the intervention period the guidelines on cholesterol testing were updated nationally. That may have been one of the reasons for the decrease in numbers of cardiovascular tests being only marginally significant. Inter-doctor variation in numbers of tests ordered decreased in both arms, but more in the intervention arm. The small group quality improvement meetings successfully discussed the transparent test ordering data and the national guidelines. Personal interaction and mutual influencing of colleagues actually occurred, implicitly resulting in an individual or group contract.[21,23]

The role of the medical coordinators as opinion-leaders also seems a crucial element of the strategy.[20,24]Questions can therefore be raised about the impact of written feedback reports in general, if these are not integrated in a wider system of quality improvement. That may have been the reason why Eccles and colleagues did not find any effect in their trial on feedback on test ordering.[9]

Some methodological comments may be made on our study. Despite the large numbers of participating PCPs a difference in baseline performance was found. It is probably due to chance as the number of randomization objects was small (n=27). Despite the lower mean number of tests at baseline the intervention arm succeed to decrease substantially. Surprisingly, the region appeared to be an important determinant in PCPs' test ordering behavior, and this finding certainly needs further investigation. We did not include a non-intervention control group, since we did not consider this as a relevant contrasting strategy. Feedback is now a regularly used strategy in primary care in the Netherlands. Unfortunately, we could not use clinical data, but since the evidence-based guidelines recommend a reduction in the total numbers of tests, the decrease we found can be interpreted as a quality improvement. Moreover, there is empirical evidence that a general reduction in test use in primary care does not lead to more referrals or substitution of care.[25,26] We expect that these limitations have had only minor impact on the results, and these results may yield two important conclusions. The first is that this new strategy can be a powerful innovative instrument to change PCPs' test ordering behavior. The strategy gives PCPs the opportunity to discuss their test ordering performance with colleagues on the basis of actual performance data, making discussions less non-committal. Our strategy also seems worthwhile because small group quality meetings can help to build up a local PCP group focusing on quality improvement. Many

test ordering problems that PCPs encounter in everyday practice, such as demands for tests by patients and changing guidelines, can be discussed and may be solved in an open and respectful discussion among professionals. Secondly, merely sending feedback reports to PCPs without extra activities, such as peer discussion or other strategies that fit in with everyday practice, does not have much impact. More effort is needed and feedback reports must fit in with a more ambitious continuous quality improvement program. Further, although our method was applied for test ordering behavior, it also seems applicable to quality improvement in other issues such as prescribing and referral behavior, and for other teams of collaborating physicians. Nation-wide implementation of this new and innovative strategy would be a logical next step and is now being prepared in the Netherlands.

References

1. Leurquin P, Van Casteren V, De Maeseneer J. Use of blood tests in general practice: a collaborative study in eight European countries. Eurosentinel Study Group. *Br J Gen Pract*. 1995;45(390):21-5.
2. Kassiser JP. Our stubborn quest for diagnostic certainty. A cause of excessive testing. *New England Journal of Medicine*. 1989;320:1489-1491.
3. Ayanian JZ, Berwick DM. Do physicians have a bias toward action? A classic study revisited. *Medical Decision Making*. 1991;11:154-158.
4. Mugford M, Banfield P, O'Hanlon M. Effects of feedback of information on clinical practice: a review. *BMJ*. 1991;303:398-402.
5. Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician performance. A systematic review of the effect of continuing medical education strategies. *JAMA*. 1995;274(9):700-5.
6. Thomson O'Brien MA, Oxman AD, Davis DA, Haynes RB, Freemantle N, Harvey EL. Audit and feedback: effects on professional practice and health care outcomes. *Cochrane Library*. 1997.
7. Thomson O'Brien MA, Oxman AD, Davis DA, Haynes RB, Freemantle N, Harvey EL. Audit and feedback versus alternative strategies: effects on professional practice and health care outcomes. *Cochrane Library*. 1997.
8. O'Connell DL, Henry D, Tomlins R. Randomised controlled trial of effect of feedback on general practitioners' prescribing in Australia. *BMJ*. 1999;318(7182):507-11.
9. Eccles M, Steen N, Grimshaw J, et al. Effect of audit and feedback, and reminder messages on primary-care radiology referrals: a randomised trial. *Lancet*. 2001;357:1406-09.
10. Wensing M, Grol R. Single and combined strategies for implementing changes in primary care: A literature review. *Int J Health Care*. 1994;6:115-32.
11. Wensing M, Van der Weijden T, Grol R. Implementing guidelines and innovations in general practice: which interventions are effective? *Br J Gen Pract*. 1998;48:991-997.
12. Grimshaw JM, Shirran L, Thomas R, et al. Changing provider behavior: an overview of systematic reviews of interventions. *Med Care*. 2001;39(8 Suppl 2):Ii2-45.
13. Winkens RAG, Pop P, Bugter-Maessen AMA, et al. Randomised controlled trial of routine individual feedback to improve rationality and reduce numbers of test requests. *Lancet*. 1995:345-502.
14. Tierney WM. Feedback of performance and diagnostic testing: lessons from Maastricht [editorial; comment]. *Med Decis Making*. 1996;16(4):418-9.
15. Grol R. Peer review in primary care. *Quality Assurance Health Care*. 1990;2:219-26.
16. Grol R, Jones R. Lessons from 20 years of implementation research. *Family Practice*. 2000;17:S32-5.
17. Brady WJ, Hissa DC, McConnell M, Wones RG. Should physicians perform their own quality assurance audits? *Journal of General Internal Medicine*. 1988;3(6):560-5.
18. Mittman B, Tonesk X, Jacobson P. Implementing clinical practice guidelines: social influence strategies and practitioner behavior change. *Quality Review Bulletin*. 1992;18:413-422.
19. Lomas J, Enkin M, Anderson GM, Hannah WJ, Vayda E, Singer J. Opinion leaders vs audit and feedback to implement practice guidelines. Delivery after previous cesarean section. *JAMA*. 1991;265(17):2202-7.
20. Thomson O'Brien MA, Oxman AD, Haynes RB, Davis DA, Freemantle N, Harvey EL. Local opinion leaders: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev*. 2000(2):Cd000125.
21. Tausch BD, Harter MC. Perceived effectiveness of diagnostic and therapeutic guidelines in primary care quality circles. *Int J Qual Health Care*. 2001;13(3):239-46.
22. Verstappen WH, Van Der Weijden T, Sijbrandij J, et al. Effect of a practice-based strategy on test ordering performance of primary care physicians: a randomized trial. *JAMA*. 2003;289(18):2407-12.
23. Spooner A, Chapple A, Roland M. What makes British general practitioners take part in a quality improvement scheme? *J Health Serv Res Policy*. 2001;6(3):145-50.
24. Borbas C, Morris N, McLaughlin B, et al. The role of clinical opinion leaders in guideline implementation and quality improvement. *Chest*. 2000;118(2 Suppl):24s-32s.
25. Winkens RA, Grol RP, Beusmans GH, Kester AD, Knottnerus JA, Pop P. Does a reduction in general practitioners' use of diagnostic tests lead to more hospital referrals? *British Journal of General Practice*. 1995;45(395):289-92.
26. Kaag ME, Wijkel D, de Jong D. Primary health care replacing hospital care --the effect on quality of care. *Int J Qual Health Care*. 1996;8:367-73.

CHAPTER VI

Comparing cost effects of two quality strategies to improve test ordering in primary care. A randomized trial.

Wim HJM Verstappen
Frits van Merode
Jeremy Grimshaw
Willy I Dubois (†)
Richard PTM Grol
Trudy van der Weijden

Published in the International Journal for Quality in Health Care. 2004;16 (In Press)

Abstract

Objective

To determine the costs and cost reductions of an innovative strategy aimed at improving test ordering routines of primary care physicians (PCPs), compared with a traditional strategy.

Design

Multicenter randomized controlled trial with randomization at local PCP group level.

Setting

Primary care, local PCP groups in 5 regions in the Netherlands with diagnostic centers.

Study participants

27 existing local PCP groups, including 194 PCPs

Intervention

The test ordering strategy was systematically developed and combined feedback, education on guidelines and quality improvement sessions in small groups. In regular quality meetings in local groups PCPs discussed each others' test ordering behavior, related it to guidelines and made individual and / or group plans for change. Thirteen groups engaged in the entire strategy (intervention arm), 14 groups received feedback only (feedback arm).

Main outcome measure

Running costs, development costs, and research costs were calculated for the intervention period per PCP per six months. The mean costs of tests ordered per PCP per six months were assessed at baseline and follow-up.

Results

The new strategy was found to cost € 702.00, the feedback strategy € 58.00. When including running costs only the intervention was found to cost € 554.70, compared to € 17.10 per PCP per six months in the feedback arm. When excluding opportunity costs for the PCPs' time spent, the intervention was found to cost € 92.70 per PCP per six months in the intervention arm. The mean costs reduction that PCPs in the intervention arm achieved by reducing unnecessary tests was € 144 larger per PCP per six months, than the PCPs in the feedback arm. (p=0.048).

Conclusion

On the basis of our findings, including the expected non-monetary benefits, we recommend further long-term effect and cost effect studies on the implementation of the quality strategy.

Key words

Quality assurance, health care; costs and costs analysis; diagnostic tests, routine; feedback

Introduction

In times of limited resources for health care, it is necessary to evaluate not only the cost-effectiveness of new treatments or procedures for patient care, but also the cost-effectiveness of new strategies to improve the quality of health care delivery. Economic evaluations of interventions aimed at changing primary care physicians' (PCPs') behavior assess the balance between benefits attained and resources needed.[1-4] Many strategies have been developed to improve PCPs' test ordering behavior, because the numbers of tests ordered by PCPs are growing in many countries, even though established guidelines regard many of these tests as unnecessary.[5,6] Rigorous studies of the effects of strategies such as educational materials, reminders, feedback, small group quality meetings and financial incentives have so far produced heterogeneous results.[7-9] A few studies investigating costs have also yielded contradictory outcomes.[10-15] We initiated an economic evaluation study to evaluate the costs and the effects of a strategy, which combines a traditional feedback strategy with a multifaceted strategy including feedback, dissemination of and group education on evidence-based guidelines, and small group quality improvement meetings in a local PCP group, using social influence as an important motivator for change.[16,17] A genuine effect of this innovative, multi-faceted strategy has been observed and presented elsewhere.[18]

The present paper provides a method for cost analyses of such quality improvement strategies, and compares the costs and cost reductions of the new strategy with one of its elements, 'classic' feedback, to assess whether implementation of the innovative test ordering quality strategy on a national scale would be worthwhile, depending not only on its effectiveness but also on the costs involved and the savings achieved.

Methods

Setting

The strategy was applied in five regions in the Netherlands with a diagnostic center, which is an institute, usually associated with a hospital, where PCPs can order tests without referring patients to the hospital. Our strategy aimed at local PCP groups, an existing infrastructure of Dutch PCPs collaborating in a specific region. These groups share patient care outside office hours and many of them also engage as a group in small group quality improvement activities, e.g. prescription quality circles. Local PCP groups with a link to one of these five diagnostic centers were eligible for the study. The medical coordinator of the diagnostic center provided the test ordering data needed, distributed the feedback reports and supervised the small group quality improvement meetings.

Design and measurements

The new strategy was tested in a multicenter randomized controlled trial. Numbers of tests ordered were assessed over a period of six months before the intervention (the baseline period) and a period of six months after the intervention (the follow-up period). The six months-intervention took place in 1999. Participating local PCP groups were randomized centrally, stratified by the size of the local PCP group and the region in order to spread the workload of the

medical coordinators of the diagnostic center. The intervention groups (intervention arm) received feedback and guidelines, and attended small group quality meetings, while the control arm groups only received feedback (feedback arm).

Intervention

The intervention consisted of the following elements: a graphical feedback report including a comparison of personal test ordering data with those of colleagues, dissemination of and group education on national, evidence-based guidelines and quality improvement meetings in small groups. During the intervention period the participating PCPs received by mail three feedback reports on the three clinical problems, together with concise information on the evidence-based clinical test ordering guidelines for these specific clinical subjects, as developed by the Dutch College of Primary Care Physicians. **Table 1** describes the clinical problems and the associated laboratory, imaging and function tests that were included in the experiment. Each report was followed by a standardized small group quality improvement meeting, at which the feedback data relating to one of the clinical problems and the guidelines were discussed. At the end of the session concrete plans for change, both at individual and local PCP group level, were established.

Effect measures and measuring instruments

Measuring costs

All costs of producing the feedback reports and organizing the small group quality meetings were calculated.

Costs were divided into the following categories:

1. Running costs

1.1. Costs of the feedback reports. Staff members of the diagnostic centers extracted and edited the data. The production costs partially depended on the number of PCPs who participated; more PCPs meant more written reports, and hence more production time and more postage costs. Secretarial time and paper costs were calculated per feedback report.

1.2. Costs of the quality meetings. Secretarial time spent for organizing the meeting and the time spent by the medical coordinator preparing and chairing the sessions were calculated per meeting per PCP.

1.3. Since each meeting lasted 1.5 hours, and we assumed half an hour for preparation and traveling, one meeting took 2 hours of the PCPs' time. PCP fees were derived from the Dutch Government's annual care review. Total national expenditure for curative PCP care in 1998 was € 1,023,227,100, which corresponds to an hourly rate of € 77. These costs were opportunity costs; in the time the PCP attended the meetings, he could not 'produce' other work.

2. Development costs

These costs covered activities for the continuation of the project, e.g. administration, organization, the development and updating of concise guideline information. A software company developed software for the production of the feedback reports, and their costs were included as well.

3. Research costs

Scientific development of the strategy, expert meetings, the financial compensation PCPs received for participating in this study with related activities, e.g. completing evaluation forms, and working up the questionnaires and evaluation forms were counted as research activities with related costs.

Registration forms measuring the time needed to extract data and to produce and send feedback reports were completed by the staff members. Costs were then calculated on the basis of the salary scales of staff members at the diagnostic center and the research department.

TABLE 1 TESTS AND COSTS OF TESTS (€) INCLUDED IN THE TRIAL

CLINICAL PROBLEM	TESTS	COSTS
	Order	9.17
Cardiovascular topics	Cholesterol	1.20
	HDL-Cholesterol	1.20
	Triglycerides	1.61
	Sodium	1.20
	Potassium	1.20
	Creatinine	1.20
	Blood urea nitrogen	1.20
	Electrocardiogram	11.36
	Exercise electrocardiogram	72.72
Upper abdominal complaints	Bilirubin	1.20
	Amylase	1.20
	Serum gluthamic-pyruvate transmaninase	1.20
	Serum gluthamic- oxaloacetic transmaninase	1.20
	Lactic dehydrogenase	1.20
	Alkaline phosphatase	1.20
	γ-Glutamyltransferase	1.20
	Ultrasound of the hepatobiliairy tract	36.36
Lower abdominal complaints	Prostate specific antigen	7.12
	X-ray abdomen	31.82
	Double contrast barium enema	86.36

Measuring cost reductions

Cost reductions were calculated using existing standard tariffs per test (**Table 1**). In the Netherlands, costs of laboratory tests are reimbursed according to standard prices for tests and orders. Reimbursement for imaging and function tests includes hospital costs and specialists' fee. Costs reductions were determined by assessing the mean difference in the costs of tests ordered per PCP and per six months between the follow-up period and the baseline period, and comparing this difference between the two arms. Cost reductions of laboratory tests were analyzed separately, because although they are a minor part of the cost reductions, they constitute the great majority of tests.

Consistent with the national, evidence-based guidelines for test ordering for the three clinical problems included in the study, a decrease in the numbers of tests was considered to represent improved patient care.

Analysis

Costs of the intervention and the feedback strategy were calculated per PCP per six months. Since the unit of randomization was the local PCP group, the unit of analysis also had to be the local PCP group. Therefore, multilevel analyses were applied to evaluate whether the local PCP groups were important determinants of the effects of the intervention. A three-level model was used with the PCP group as level 3, the PCPs as level 2, and the numbers of tests as level 1. This model was analyzed using SAS PROC MIXED. Multilevel baseline analyses showed that analyses could be performed without the local PCP groups. All effects were analyzed with analyses of covariance using the costs of tests during the follow-up period as the dependent variable and the costs of tests at baseline and the region, which appeared to be an important determinant, as independent variables. A sensitivity analysis was performed by varying the inclusion of the various cost categories and cost reductions.

Results

A total of 38 local PCP groups were invited by open recruitment to take part in this trial. Twenty-seven local groups with 194 PCPs immediately expressed their willingness to participate. After randomization, the intervention arm included 13 local PCP groups and the feedback arm 14. **Figure 1** describes the study design and shows that follow-up data were unavailable for 20 PCPs. **Table 2** shows that there were no differences in individual PCP characteristics between the two arms. There was a large, but statistically not significant, difference in costs of laboratory and all tests ordered per PCP between the two arms at baseline.

Costs of the strategy and cost reductions in test ordering

Table 3 shows the total costs of the intervention. Concerning the running costs of the strategy, the cost of one feedback report per PCP was € 5.70. The costs per PCP per quality meeting were € 25.20 for 4.25 hours of coordination time, including secretarial, preparation, meeting and traveling time. The opportunity costs of the PCPs' time spent attending the meetings were 2 hours x € 77 = € 154 per PCP per meeting. About the development costs, guidelines were only used in

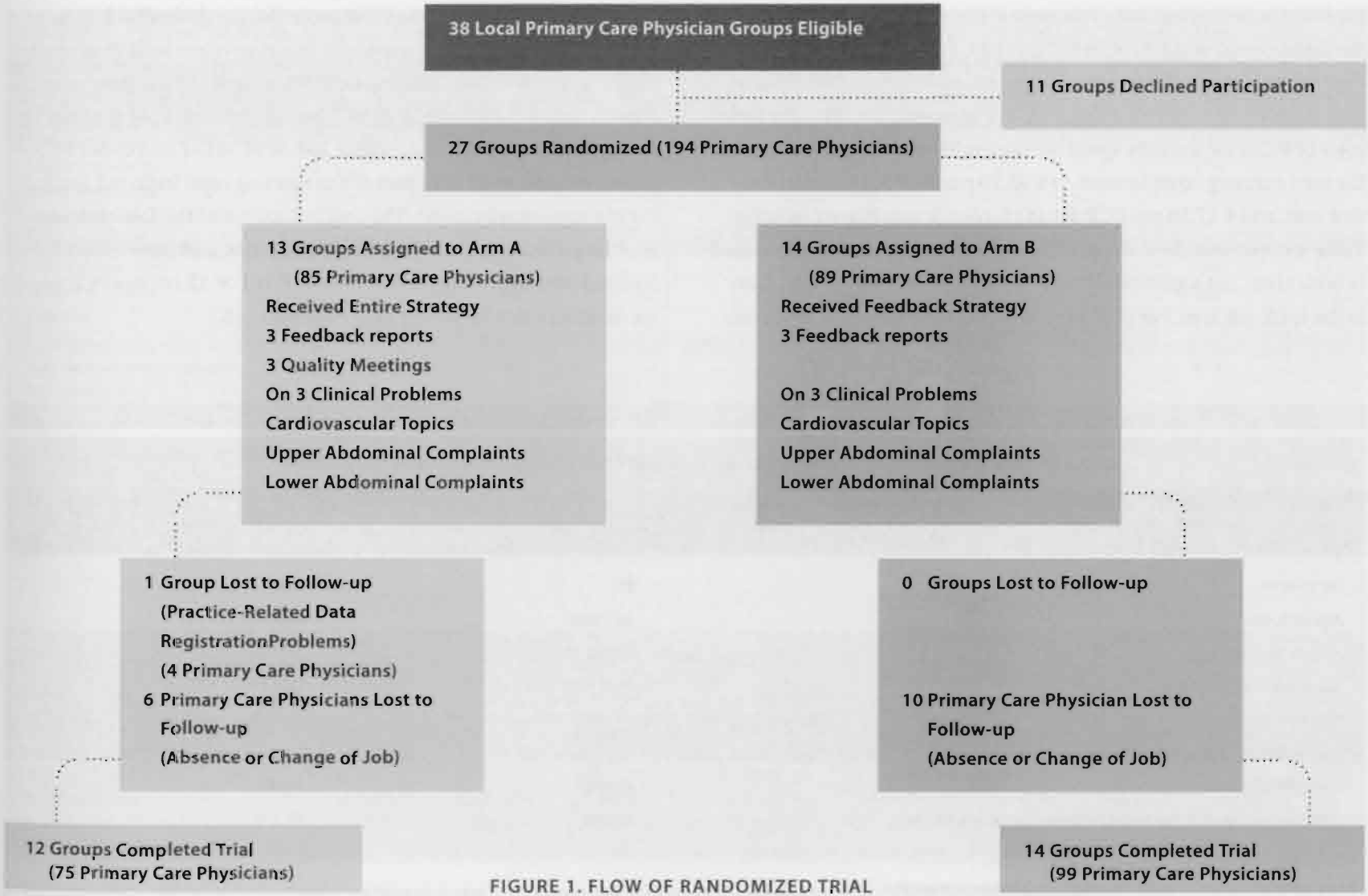


FIGURE 1. FLOW OF RANDOMIZED TRIAL

the intervention group and costs were € 4 per PCP. The total cost of the intervention was € 65,998: € 702 per PCP per six months for the intervention arm and € 58 per PCP per six months for the feedback arm. If only part of the running costs are counted (the opportunity costs of PCPs for the time spent for the quality meetings are excluded) the total running costs amount to € 92.70 per PCP for the intervention arm, and € 17.10 per PCP for the feedback arm per six months. **Table 4** shows that the costs of laboratory as well as all tests, decreased in both arms, but significantly more so in the intervention arm than in the feedback arm. Per PCP per six months the total cost reduction in the intervention arm was € 144 more than in de feedback arm. **Table 5** shows results of a sensitivity analysis. When including opportunity costs for PCPs' attending time, the costs for the intervention arm exceeded the cost reductions. The cost reductions of the intervention arm exceeded the costs with € 208.30 (€ 301- € 92.70) per PCP per six months, with only part of the running costs included (excluding the opportunity costs). The cost reductions of the feedback arm were larger than its costs for all cost categories, and introducing the feedback strategy would save € 143.90 (€ 161- € 17.10) per PCP per six months, when including only running costs.

TABLE 2 STUDY POPULATION CHARACTERISTICS AT INDIVIDUAL PRIMARY CARE PHYSICIAN LEVEL

	INTERVENTION ARM	FEEDBACK ARM
No. of PCPs	85	109
Age (SD), year	46.2 (6.6)	46.2 (6.6)
Female, No. (%)	14 (16)	11 (10)
No. of patients per physician, mean (SD) *	2587 (641)	2444 (416)
Patients older than 65y, % mean (SD)	15 (6.8)	15 (6.5)
Physician with a part-time working factor, % mean (SD)	91 (15)	92 (12)
Physician with a solo practice, No. (%)	43 (51)	44 (40)
Physician who uses computerized registration system, No.(%)	66 (78)	75 (69)

* Total practice population for whom the primary care physician is responsible.

TABLE 3 INTERVENTION COSTS FOR THE STUDY POPULATION AND PER PRIMARY CARE PHYSICIAN IN EACH ARM IN € PER 6 MONTHS.

TYPE OF COSTS	COSTS TOTAL INTERVENTION		COSTS PER PCP INTERVENTION ARM (N= 85)		COSTS PER PCP FEEDBACK ARM (N= 109)	
Running costs	49.014		554.70		17.10	
Feedback reports (3x)		3.317		17.10		17.10
Quality meetings (3x)		6.427		75.60		
Opportunity costs*		39.270		462		
Development costs	3.861		22.40		18.00	
Continuation activities		2.484		12.80		12.80
Software development		1.000**		5.20		5.20
Guidelines		377**		4.40		
Research costs	13.123		124.90		22.90	
Scientific development/evaluation		4.453		22.90		22.90
PCP compensation		8.670		102		
TOTAL COSTS	65.998		702.00		58.00	

* Based on hourly fees for curative PCP care in 1998 as derived from the Dutch Government's annual care review

** Discounting period 5 years

TABLE 4 EFFECTS OF THE STRATEGY BY ANALYSIS OF COVARIANCE ADJUSTED FOR COSTS OF NUMBERS OF TESTS AT BASELINE AND FOR THE REGION ON THE MEAN COSTS IN € (SD) OF LABORATORY AND ALL TESTS ORDERED PER PRIMARY CARE PHYSICIAN PER 6 MONTHS

	INTERVENTION ARM (N= 75)		FEEDBACK ARM (N=99)		β*	S.E. B	P	95% CI.
	Baseline	Follow-up	Baseline	Follow-up				
Total costs laboratory tests	596 (407)	517 (313)	656 (437)	633 (393)	-64	66	.0027	-106; -23
Total costs all tests	1541(1023)	1240(720)	1763 (1268)	1602 (1016)	-144	72	.048	-287; -2

β= intervention effect = total change between baseline and follow-up in mean costs of tests in the intervention group – total change in mean costs of tests in the control group

TABLE 5 COSTS AND COST REDUCTIONS (€) PER PCP PER SIX MONTHS.

COSTS	INTERVENTION ARM	FEEDBACK ARM
All costs[1]	701.00	58.00
Only running costs[2]	554.70	17.10
Running costs, no opportunity costs[3]	92.70	17.10
Cost reductions[4]	**301.00**	**161.00**

[1]All costs: include running costs, development costs and research costs of the strategy.

[2]Running costs: include costs of the feedback reports, small group quality meetings and opportunity costs.

[3]Opportunity costs: costs of the primary care physicians' time spent attending the small group quality meetings. One meeting took 2 hours of the primary care physicians' time (including preparation and traveling time). A primary care physician hourly rate of € 77 was derived from Dutch Goverment's annual review.

[4]Cost reductions were differences in costs of test at follow-up and at baseline, and were calculated using existing standard tariffs per test.

Discussion

The present paper evaluates costs and cost reductions of an innovative strategy to improve PCPs' test ordering, involving feedback, education on guidelines, peer interaction and social influence, by comparing it with a traditional approach involving only the provision of feedback. The new strategy improved test ordering more substantially and consistently, and, besides the favorable clinical effects, appears to bring about more cost reduction than feedback alone when not counting the opportunity costs[18-20]. Introducing this effective strategy in the Netherlands, with its about 7100 PCPs practising, would then save € 1,478,930 (7100 x € 208.30) in the first six months.

There are some methodological aspects of our study that need to be considered. Concerning cost reductions, a reduction in the number of laboratory tests ordered does not always influence laboratory costs; for example, a diagnostic apparatus performing fewer tests costs the same amount of money and only a large reduction can mean that fewer laboratory staff are needed. We could not include such potential cost reductions. For instance, not performing a redundant test also implies that a patient does not have to take time off work. More importantly, we were unable to assess the cost reductions achieved by not performing tests that would result in false-positive findings. Such test results may lead to a cascade of further testing, or inappropriate treatment or referrals, and as a result of better diagnosing patients costs are saved in the long run. The limited time frame of the study prevented us to study these effects, since patients included in our study should be monitored for several years. For the same reason we were unable to assess possible learning effects, which could mean that quality activities may become less time-consuming over time even if the approach is directed to other clinical problems. Finally, under use of tests is another possible danger that was not assessed.

Our study deals with some interesting and important topics for costs analyses of quality improvement studies. As in many quality improvement studies only intermediate effect measures instead of patient outcome measures were available. Since negative effects on patient' outcome are not expected in the quality strategy, these kind of cost analyses can be seen as cost minimization analyses.[21] The analyses were done from a societal perspective, but the perspective of the physicians involved may also be important. Further, we focused on the costs and cost reductions, expressed in monetary units, but with our new strategy we may expect also non-monetary benefits related to the strategy, such as improvement of the PCPs' clinical knowledge and job satisfaction, and, of course, it is difficult to quantify these important benefits in such a cost analysis. There is some empirical evidence that participating in such quality improvement activities may increase PCPs' job satisfaction.[22,23] Moreover, we calculated the opportunity costs for the time spent by PCPs in attending the quality meetings. In general, these opportunity costs, should be included because they weigh (in monetary units) the time needed for conducting the activities considered in this study and not available anymore for other activities. Which is the reason they are named opportunity costs. However, it remains debatable if these opportunity costs have to be accounted for in the Netherlands because here PCPs are obliged to engage in continuous medical education programs, such as our quality strategy, up to 40 hours a year, and health insurers partially include compulsory continuous medical education in the national tariffs. Furthermore, it was found to be difficult to differentiate between development and research costs, and we decided to define only the costs of the expert meetings and the scientific effect evaluations, including the compensation for the research activities of participating PCPs, as research costs. Nevertheless, it is debatable whether these costs have to be accounted for, and researchers have to explicit their choice. In costs analyses research costs usually are excluded, and it is debatable how to handle development costs, because some of these costs will be necessary when implementing a strategy at a broader scale. Concluding, we evaluated costs and cost reductions of our strategy without counting the scientific and development costs. However, including the development costs in our sensitivity analysis did not change our results.

Surprisingly, cost effects have usually not been evaluated in quality improvement studies, perhaps because, as was mentioned above, many problems can be expected.[24] Recently, Mason et al. provided a framework for exploring the economics of improving quality of care by means of influencing physicians' behavior, using clinical outcome data at patient level.[25] The present paper provides a method for cost analyses of quality improvement strategies, where it is difficult or even impossible to perform a real cost-effectiveness study because of lack of clinical data.

In conclusion, in the Dutch situation the innovative test ordering strategy reveals considerable cost reductions in the first six months when not counting the opportunity costs for the time spent by PCPs. Because, contrary to the feedback strategy, also non-monetary benefits can be expected, we suggest that PCPs organizations stimulate local PCP groups to participate in this new strategy.

References

1. Mason, J., et al., A framework for incorporating cost-effectiveness in evidence-based clinical practice guidelines. Health policy, 1999. 47(1): p. 37-52.
2. Sculpher, M., Evaluating the cost-effectiveness of interventions designed to increase the utilization of evidence-based guidelines. Fam Pract, 2000. 17(31): p. 1s26-31.
3. Mauskopf, J.A., et al., The role of cost-consequence analysis in healthcare decision-making. PharmacoEconomics, 1998. 13(3): p. 277-88.
4. McIntosh, E., C. Donaldson, and M. Ryan, Recent advances in the methods of cost-benefit analysis in healthcare. Matching the art to the science. PharmacoEconomics, 1999. 15(4): p. 357-67.
5. Leurquin, P., V. Van Casteren, and J. De Maeseneer, Use of blood tests in general practice: a collaborative study in eight European countries. Eurosentinel Study Group. Br J Gen Pract, 1995. 45(390): p. 21-5.
6. Kristiansen, I.S. and P. Hjortdahl, The general practitioner and laboratory utilization: why does it vary? Fam Pract, 1992. 9: p. 22-7.
7. Oxman, A.D., et al., No magic bullets: A systematic review of 102 trials of interventions to improve professional practice. Can Med Ass J, 1995. 153: p. 1423-1431.
8. Bero, L.A., et al., Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. The Cochrane Effective Practice and Organization of Care Review Group. BMJ, 1998. 317(7156): p. 465-8.
9. Solomon, D.H., et al., Techniques to improve physicians' use of diagnostic tests. A new conceptual framework. JAMA, 1998. 280: p. 2020-2027.
10. Etter, J.F. and T.V. Perneger, Health care expenditures after introduction of a gatekeeper and a global budget in a Swiss health insurance plan. Journal of Epidemiology & Community Health, 1998. 52(6): p. 370-376.
11. Helgesen, F.A., Follow-up of prostate cancer patients by on-demand contacts with a specialist nurse: a randomized study, in: Scand-J-Urol-Nephrol. 2000 Feb; 34(1): 55-61. 2000.
12. Larsson, A., et al., Effects of an education programme to change clinical laboratory testing habits in primary care. Scand J Prim Health Care, 1999. 17: p. 238-243.
13. Wensing, M. and R. Grol, Single and combined strategies for implementing changes in primary care: A literature review. Int J Health Care, 1994. 6: p. 115-32.
14. Wensing, M., T. Van der Weijden, and R. Grol, Implementing guidelines and innovations in general practice: which interventions are effective? Br J Gen Pract, 1998. 48: p. 991-997.
15. Thomson O'Brien, M.A., et al., Audit and feedback versus alternative strategies: effects on professional practice and health care outcomes. Cochrane Library, 1997.
16. Mittman, B., X. Tonesk, and P. Jacobson, Implementing clinical practice guidelines: social influence strategies and practitioner behavior change. Quality Review Bulletin, 1992. 18: p. 413-422.
17. Grol, R., Successes and failures in the implementation of evidence-based guidelines for clinical practice. Med Care, 2001. 39(8 Suppl 2): p. Ii46-54.
18. Verstappen, W.H., et al., Effect of a practice-based strategy on test ordering performance of primary care physicians: a randomized trial. JAMA, 2003. 289(18): p. 2407-12.
19. Little, P., et al., Why do GPs perform investigations?: The medical and social agendas in arranging back X-rays. Fam Pract, 1998. 15(3): p. 264-5.
20. McDonald, I.G., et al., Opening Pandora's box: the unpredictability of reassurance by a normal test result [see comments]. BMJ, 1996. 313(7053): p. 329-32.
21. Buxton, M.J., et al., Modelling in economic evaluation: an unavoidable fact of life. Health economics, 1997. 6(3): p. 217-27.
22. Tausch, B.D. and M.C. Harter, Perceived effectiveness of diagnostic and therapeutic guidelines in primary care quality circles. Int J Qual Health Care, 2001. 13(3): p. 239-46.
23. Spooner, A., A. Chapple, and M. Roland, What makes British general practitioners take part in a quality improvement scheme? J Health Serv Res Policy, 2001. 6(3): p. 145-50.
24. Brown, C.A., C.R. Belfield, and S.J. Field, Cost effectiveness of continuing professional development in health care: a critical review of the evidence. BMJ, 2002. 324(7338): p. 652-5.
25. Mason, J., et al., When is it cost-effective to change the behavior of health professionals? JAMA, 2001. 286(23): p. 2988-92.

CHAPTER VII

Lessons learnt from applying an innovative, small group quality improvement strategy on test ordering in general practice.

Wim HJM Verstappen
Trudy van der Weijden
Willy I Dubois (†)
Ivo JM Smeele
Marianne A Meulepas
Richard PTM Grol

Published in Quality in Primary Care 2004;12: 79-85

Abstract

Objective

Evaluation of the feasibility of an innovative strategy to improve GPs' test ordering behaviour, and to further improve continuous professional development.

Design

Prospective process evaluation of the use and appraisal of the strategy during the first and second years of a trial.

Setting

General practice, local GP groups, diagnostic centres.

Intervention

The new strategy combines written feedback, education on clinical guidelines and continuous quality improvement sessions, quality circles, in small local GP groups. An important feature of the written feedback was a comparison of the behaviour of individual GPs with that of their colleagues. Mutual feedback by working in pairs, discussion on national guidelines, and making plans for change were important features of the group sessions. The strategy has an iterative character.

Results

All 194 participating GPs received the planned six feedback reports. Data from 156 meetings of 26 local GP groups showed a participation rate of 81% (95% CI: 77%-85%) in the first year and 73% (95% CI: 68%-77%) in the second. Meetings included mutual feedback by working in pairs (used in 73% of the sessions in the first year and 61% in the second year), individual plans for change (96% in the first year, 92% in the second year) and group plans for change (71% in the first year, 54% in the second year). In the first year GPs expressed their level of satisfaction with the approach in a score of 7.55 on a scale of 0 - 10 (95% CI 7.46-7.64); average score in the second year was 7.51 (95% CI 7.30-7.74).

Conclusion

The innovative test ordering strategy seems a feasible tool for continuous improvement of GPs' test ordering behaviour, fitting in well with local and regional quality improvement efforts for isolated working GPs.

Key words

quality assurance, health care; evaluation studies, primary health care, professional practice, test ordering, feedback, guidelines.

Introduction

Numbers of tests ordered by general practitioners (GPs) is growing, and inter-doctor variation is shown to be large.[1-3] It is as yet unclear, however, what would be the best method to influence GPs' test ordering behaviour.[4] Studies evaluating different types of interventions and strategies for this purpose have, so far, produced heterogeneous results[1-5]. No particular type of intervention was found to be inherently effective; multifaceted approaches have proved to be superior to single methods in some analyses, but not in other.[6,7] Audit and feedback were found to be effective in specific settings[8,9], while written, personal feedback on test ordering by peers or opinion leaders has also been found to improve test ordering behaviour.[10] It seems particularly important in this respect to make use of interventions in addition to professionally oriented interventions, because the success rates of particular strategies seem to be highly dependent on the extent to which they fit in with the local context and the practitioners' daily work routine.[11] A multi-faceted strategy combining comparative feedback on tests ordered, group education on guidelines, and small group quality improvement meetings in a local GP group, with social influence as an important motivator for change, was expected to offer good prospects.[12,13] The strategy also fits in well with the work setting of many GPs in European and non-European countries, which are often characterised by small practices, relatively isolated settings and a desire for more contacts with peers.

The favourable clinical effects of this strategy were reported elsewhere.[14] Nowadays process evaluations of quality improvement strategies are seen as a necessary addition to effect studies to learn about important elements of change.[15] It was therefore important to determine to what extent the intended elements of the multifaceted strategy were accepted and actually used by the participants and to assess their opinion on the key elements of the feedback and interactive quality circles between colleagues.[16,17] The present paper focuses on the feasibility of this innovative strategy in view of a possible implementation at a larger scale, and it also assesses important elements from the perspective of further improving continuous professional development (CPD) of general practitioners.

Methods

Design and subjects

Between January 1999 and October 2000, the new strategy was evaluated in five regions in the Netherlands, and a process evaluation was done prospectively. Coordination of the feedback and supervision of the group meetings was provided by the five diagnostic centres, which are a special facility where GPs can order laboratory, imaging and function tests without referring patients for specialist care. One of the tasks of the medical coordinator of these centres is to give feedback to GPs on their test ordering behaviour.

Local GP groups that referred their patients to one of the five participating diagnostic centres were invited to take part in the study. Local GP groups are an existing part of the infrastructure of Dutch GPs collaborating in a specific region. One of their tasks is to organise care during out of office hours, while CPD is another important activity in many of these local groups.

Intervention: the improvement strategy

The intervention consisted of the following elements: personalised graphical feedback, including a comparison of each GP's own test ordering data with those of colleagues, guideline dissemination and continuous quality improvement meetings in small groups, organised and chaired by the medical coordinator of the diagnostic centre. The strategy was patient care oriented rather than test oriented, in that it did not focus on the volume of specific tests, but on specific clinical problems and associated laboratory, imaging and function tests relevant to daily GP practice (**Table 1**). GPs received three different feedback reports per year on three different clinical problems, together with the national, evidencebased guidelines on test ordering of these specific clinical subjects. This was followed by 90-minute structured meeting two weeks later, at which one of the clinical problems was discussed. The small group meetings or quality circles consisted of three major components. The first was mutual personal feedback by peers, who worked in pairs at the start of the meeting. This was assumed to be a safe method of peer review. The second component was an interactive

TABLE 1 CLINICAL PROBLEMS AND ASSOCIATED TESTS USED IN THE TRIAL

	CLINICAL PROBLEMS / TESTS		CLINICAL PROBLEMS / TESTS
A1	**Cardiovascular topics** Cholesterol, subfractions, potassium, sodium, serum creatinine, blood urea nitrogen, (exercise) EKG	B1	**COPD/Asthma** Pulmonary function test, allergic screening test, immunoglobulin E, chest X-ray
A2	**Upper abdominal complaints** Alanine aminotransferine, aspartate aminotransferase, lactic dehydrogenase, amylase, γ-glutamyltransferase, bilirubin, alkalic phosphatase, ultrasound scans of hepatobiliary tract	B2	**General malaise /fatigue/ vague complaints** ESR, Haemoglobin + - indices, haematocrit, white blood count, thyreoid stimulating hormone, monospot
A3	**Lower abdominal complaints** Prostate-specific antigen, C-Reactive protein, ultrasound of the kidney, IVP, double contrast barium enema, sigmoidoscopy	B3	**Joint degeneration / complaints** ESR, uric acid, rheumatoid factors, X-rays of lumbar spine, cervical spine, shoulder, knee, hip

group education of national guidelines to be able to relate own and each others' test ordering behaviour with them. The third was the development of individual and group plans for change to stimulate GPs to really put their plans into daily practice. This schedule was repeated a year later, using the same three clinical problems, to assess whether a GP or GP group had implemented the plans for change and to initiate further improvements. This iterative aspect was another important feature of the strategy.

Variables and instruments

The feasibility of the strategy was tested by a prospective process evaluation, focusing on 6 variables: (1) the timely production and provision of the feedback reports, (2) the GPs' appreciation of the feedback, (3) the attendance at the meetings and (4) the GPs' appreciation of the meetings. These four variables were measured by means of a one-page standardised questionnaire, which was completed by the attending GPs after each meeting.

Appreciation was measured on a scale of 0-10. (5) With a checklist the medical coordinators recorded actual activities at the meetings, e.g. mutual feedback, discussions on guidelines and plans for change. (6) Finally, individual and group plans for change were drawn up by the participating GPs, written down and collected by the coordinators of two regions during the meetings.

Statistical analysis

Analyses were performed separately for the first and second year, in view of the iterative aspect of the intervention. For the same reason differences in attendance between the first and second years were tested for significance using the McNemar test for paired variables. Subgroup analyses for regions and for clinical problems were performed for some of the parameters to see if region and clinical problems were important determinants for the process evaluation. Because there were differences in group size, Spearman's correlation coefficients were calculated to see if group size was correlated with items from the actual activities questionnaire.

ANOVA and multivariate regression analyses were done on the GPs' appreciation of the feedback reports, using the clinical problem, the region and the local GP group as independent variables.

Results

A total of 37 local GP groups were invited to take part in the trial. The total study population was 193 GPs, belonging to 26 local GP groups that were willing to participate. Individual GP and GP practice characteristics were largely similar to those of the Dutch GP population as a whole, except for type of practice: two-person practices were under represented, while group practices were over represented.

The mean group size was 7.4 ± 2.7(SD), minimum 3, maximum 12. A total of 1158 (6x193) written feedback reports were sent out, and 156 small group quality improvement meetings were held. A total of 850 GP questionnaires were analysed, 455 in the first year and 395 in the second. The response by the participating GPs to the questionnaires was 97 % in the first year and 93% in the second year. The response by the medical coordinators was 100% in the first year and 99% in the second year.

TABLE 2 APPRECIATION OF WRITTEN FEEDBACK REPORTS, FILLED IN BY 193 GPS (MEAN; 95%CI; SCALE 0-10).

FIRST YEAR		MEETING 1		MEETING 2		MEETING 3	
CLINICAL PROBLEM	TOTAL	A1	B1	A2	B2	A3	B3
Appreciation of written report	7.51 (7.42-7.60)	7.69	7.51	7.59	7.25	7.61	7.45
SECOND YEAR		**MEETING 4**		**MEETING 5**		**MEETING 6**	
CLINICAL PROBLEM	TOTAL	A1	B1	B1	B2	A3	B3
Appreciation of written report	7.46 (7.37-7.56)	7.62	7.05	7.71	7.52	7.50	7.38

TABLE 3 ACTIVITIES PERFORMED DURING THE MEETINGS IN THE FIRST AND SECOND YEARS, DESCRIBED BY MEDICAL COORDINATORS. (PER CLINICAL PROBLEM, IN PERCENTAGES).

	PER CLINICAL PROBLEM						
PERFORMED ACTIVITIES FIRST YEAR	A1	A2	A3	B1	B2	B3	TOTAL (95% C.I.)
Appraisal of own behaviour	100	100	100	100	100	100	100
Pair work	62	75	64	92	92	62	73(63-83)
Discussing relation guidelines	100	100	100	100	100	100	100
Individual plans	92	92	100	92	100	100	96(92-101)
Group plans	85	50	50	50	85	69	71(60-81)
PERFORMED ACTIVITIES SECOND YEAR	A1	A2	A3	B1	B2	B3	TOTAL (95% C.I.)
Appraisal of own behaviour	100	100	100	100	100	100	100
Pair work	58	50	58	58	58	69	61(49-72)
Individual plans	92	100	100	100	92	92	92(86-98)
Group plans	50	67	45	58	58	54	54(42-65)
Discussing previously drawn up plans for change	100	100	100	100	100	100	100

Each participant received all six feedback reports as planned. It proved to be possible to produce and disseminate the feedback in time. The GPs gave a favourable assessment of the feedback reports in both years (**Table 2**). Multivariate regression analysis showed that the region where the GP practised, the local GP group and the clinical problem had no significant influence on the appreciation in the first year. In the second intervention year, the clinical problem did influence the appreciation of the report ($p=0.03$), in that the appreciation of the feedback report on COPD / asthma related tests decreased in the second year. Attendance at the meetings in the first year was on average 81% (95% CI: 0.77-0.85); in the second year attendance decreased to 73% (95% CI: 0.68-0.77) ($p < 0.05$, Mc Nemar test). Only two of the 196 GPs never visited any of the meetings. Subgroup analysis showed that there were no significant differences in attendance per region or per clinical problem ($p > 0.05$). Overall, participants expressed favourable opinions on the new strategy: the average appreciation score was 7.55 (95% CI 7.46-7.64; scale 0-10; min 4; max 10) in the first year and 7.51 (95% CI 7.38-7.65) in the second.

Table 3 describes the actual activities in the meetings during the two trial years. Discussion of participants' own test ordering behaviour was performed according to plan in all meetings. As planned, all groups discussed the relation with the evidence-based guidelines as well –in the second year- the plans for change made in the first year. In the first year, participants worked more in pairs than in the second year; in two out of the five regions less pair work was undertaken. There was a significant positive correlation of 0.38 ($p<0.01$) between a larger group size and more pair work in the first year, which disappeared in the second year. **Table 3** also shows that GPs made individual plans for change in most meetings. Most groups also made group plans for change, although this decreased in the second year.

Table 4 gives the most-mentioned individual plans for change per clinical problem. Most plans concerned a decrease in the number of tests, except for lung function tests. An example of such an individual commitment was, "I will order fewer Hb tests, because I realise that this test does not give much information in patients with vague complaints". The second year the number of individual plans decreased, except for the clinical problem general malaise/ vague complaints. Plans at group level were also made, e.g. the plan to use the same patient brochure to inform patients about the use of cholesterol tests or the arrangement to follow the national guideline on delaying testing in patients with vague complaints. All results show that the quality circles were an essential element in the improvement strategy.

Discussion

The innovative, multi-faceted strategy for improving test ordering behaviour was favourably evaluated by a large GP population. All local GP groups expressed a desire for continuation of the meetings after the experiment. The new strategy utilised peer influence among GPs, and gave GPs the opportunity to openly discuss their test ordering behaviour with colleagues.

The results may be biased, since the study population differed slightly from the Dutch GP population only regarding the type of practice.

However, there is no reason to assume that these minor differences influenced the external validity of the study. The decision to focus on clinical problems instead of tests was a good choice, since it allowed the feedback and group work to be linked to national evidence-based guidelines. GPs appreciated this approach, because it was also closely related to their everyday work routine. They stated that this type of feedback definitely had added value, because comparison with colleagues made them more conscious of their own behaviour and motivated them to change.[18 19] Their main criticism was the validity of the numbers of tests in the feedback and the absence of patient-related data.
Working in pairs to discuss the feedback report at the start of the meetings made most GPs feel safe, especially in the first year. After a while, it may become less needed, because participants may then feel more safe about discussing their own behaviour within the group as a whole. This is probably why the use of pair work decreased in the second year. Drawing up concrete individual and, if possible, group plans for change that are checked later is a crucial and innovative aspect of this strategy. Most GPs made individual plans for decreasing the numbers of certain tests.
However, lack of experience in drawing up and GPs were excited to find in the second year that they had indeed changed in accordance with their plans, and they were then usually more motivated to implement further changes. Nevertheless, individual plans for change were not always adhered to. Making group plans for change can be difficult, due to lack of confidence or lack of familiarity with entering into this kind of commitment in a GP group. However, almost two-thirds of the meetings managed to draw up group plans for change.

An explanation for the slight decrease in the attendance rate in the second year might be that the same clinical problems were discussed, with some GPs stating that they did not expect to learn anything new, and they preferred to discuss a new clinical subject at each meeting in addition to evaluating previous plans for change.

There is some empirical evidence that participating in quality circles may increase GPs' job satisfaction, and this powerful, interactive group strategy fits well within the growing need of transparent health care with positive use of actual clinical data for continuous professional development in order to further improve clinical practice.[16 20] The following lessons for the CPD of GPs can be learnt. First, GPs appreciate the combination of individual feedback, discussions about guidelines and small group quality improvement meetings driven by peer influence. A second important element is the fact that GPs are prepared to discuss personal, transparent data openly in a group of colleagues. Thirdly, another important element is the focus on daily, clinical GP problems. In our study GPs preferred to talk about clinical problems and tests linked to these problems, rather than to discuss abstract phenomena like total test ordering volume or the ordering of specific tests. Finally, the strategy must fit in with the GPs' daily practice routine and should be aimed at local collaboration in teams or groups.

Acknowledgements: The authors gratefully acknowledge the financial contribution to the study by the Dutch College for Health Insurances..

TABLE 4 INDIVIDUAL PLANS FOR CHANGE MADE BY GPS IN TWO REGIONS DURING THE TWO-YEARS PERIOD. (ONLY ITEMS MENTIONED BY AT LEAST FOUR TIMES WERE COUNTED)

CLINICAL PROBLEMS/ TESTS A TOTAL N =34 GPS*	1999	2000	CLINICAL PROBLEMS/ TESTS B TOTAL N = 37 GPS*	1999	2000
A1 CARDIOVASCULAR DISEASES/HYPERTENSION			**A2 COPD/ASTHMA**		
Decrease:			*Decrease:*		
Cholesterol	10	4	Immoglobulin E	10	
Subfractions	5	10	Allergic screening test	8	
Exercise ECG	4		Chest X-ray	6	8
			Increase:		
			Pulmonary function test	7	
B1 UPPER ABDOMINAL COMPLAINTS			**B2 GENERAL MALAISE/VAGUE COMPLAINTS**		
Decrease:			*Decrease:*		
ASAT	10	6	Leucocytes	15	16
γ-glutamyltransferase	10	10	MCV and indices	9	11
LDH	9		TSH	6	6
Alkalic phosphatase	8		HB	5	7
ALAT	6	7	ESR	4	9
Ultrasound scan of hepatobliary tract	5	5	Leucocytes differential count	4	4
Bilirubin	4	4	Mononucleosis test		7
A3 LOWER ABDOMINAL COMPLAINTS			**B3 JOINT DEGENERATION/JOINT COMPLAINTS**		
Decrease:			*Decrease:*		
Prostate specific antigen	12	8	Uric acid	14	10
CRP	11	6	Rheumatoid factors	4	4
IVP	6		X-ray of shoulder		6

* GPs were allowed to indicate more than one item.

References

1. Axt-Adam P, van der Wouden JC, van der Does E. Influencing behavior of physicians ordering laboratory tests: a literature study. *Med Care* 1993;31:784-94.
2. Buntinx F, Winkens RAG, Grol RPTM, Knotnerus JA. Influencing diagnostic and preventive performance in ambulatory care by feedback and reminders. A review. *Fam Pract* 1993;10:219-28.
3. Oxman AD, Thomson MA, Davis DA, Haynes RB. No magic bullets: A systematic review of 102 trials of interventions to improve professional practice. *Can Med Ass J* 1995;153:1423-1431.
4. Solomon DH, Hashimoto H, Daltroy L, Liang MH. Techniques to improve physicians' use of diagnostic tests. A new conceptual framework. *JAMA* 1998;280:2020-2027.
5. Mugford M, Banfield P, O'Hanlon M. Effects of feedback of information on clinical practice: a review. *BMJ* 1991;303:398-402.
6. Wensing M, Grol R. Single and combined strategies for implementing changes in primary care: A literature review. *Int J Health Care* 1994;6:115-32.
7. Wensing M, Van der Weijden T, Grol R. Implementing guidelines and innovations in general practice: which interventions are effective? *Br J Gen Pract* 1998;48:991-997.
8. Thomson O'Brien MA, Oxman AD, Davis DA, Haynes RB, Freemantle N, Harvey EL. Audit and feedback: effects on professional practice and health care outcomes. *Cochrane Library* 1997.
9. Thomson O'Brien MA, Oxman AD, Davis DA, Haynes RB, Freemantle N, Harvey EL. Audit and feedback versus alternative strategies: effects on professional practice and health care outcomes. *Cochrane Library* 1997.
10. Winkens RA, Pop P, Grol RP, Kester AD, Knottnerus JA. Effect of feedback on test ordering behaviour of general practitioners. *BMJ* 1992;304:1093-1096.
11. Van der Weijden T, Grol R, Winkens R, Buntinx F, ter Riet G, Klazinga N. Interventions aimed at influencing the use of diagnostic tests. The relevance of attention for contextual factors.[Protocol]. *Cochrane Library* 2001.
12. Grol R. Peer review in primary care. *Quality Assurance Health Care* 1990;2:219-26.
13. Mittman B, Tonesk X, Jacobson P. Implementing clinical practice guidelines: social influence strategies and practitioner behavior change. *Quality Review Bulletin* 1992;18:413-422.
14. Verstappen WH, Van Der Weijden T, Sijbrandij J, Smeele I, Hermsen J, Grimshaw J, et al. Effect of a practice-based strategy on test ordering performance of primary care physicians: a randomized trial. *JAMA* 2003;289(18):2407-12.
15. Hulscher ME, Laurant MG, Grol RP. Process evaluation on quality improvement interventions. *Qual Saf Health Care* 2003;12(1):40-6.
16. Tausch BD, Harter MC. Perceived effectiveness of diagnostic and therapeutic guidelines in primary care quality circles. *Int J Qual Health Care* 2001;13(3):239-46.
17. Szecsenyi J, Beyer M, Gerlach F, al e. The development of quality circles/peer review groups as a method of quality improvement in Europe. results of a survey in 26 European countries. *Family Practice* 2003;20:443-52.
18. Weissman NW, Allison JJ, Kiefe CI, Farmer RM, Weaver MT, Williams OD, et al. Achievable benchmarks of care: the ABCs of benchmarking. *J Eval Clin Pract* 1999;5(3):269-81.
19. 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):2871-9.
20. Spooner A, Chapple A, Roland M. What makes British general practitioners take part in a quality improvement scheme? *J Health Serv Res Policy* 2001;6(3):145-50.

CHAPTER VIII

Block design allowed for control of the Hawthorne effect in a randomised controlled trial of test ordering.

Wim HJM Verstappen
Trudy van der Weijden
Gerben ter Riet
Jeremy Grimshaw
Ron Winkens
Richard PTM Grol

Published in the Journal of Clinical Epidemiology: 2004: 57 (In Press)

Abstract

Objective

To evaluate the value of balanced incomplete block designs in quality improvement research, and their capacity to control for the Hawthorne effect: the phenomenon that the mere taking part of a professional in a trial and his or her awareness being observed influences performance.

Study design and setting

In a clustered trial, GP teams were randomised into three arms and received a quality improvement intervention on test ordering, relating to tests for two groups of clinical problems, called A tests and B tests. In the two trials within the block design we tried to control for the Hawthorne effect by comparing the complete intervention in both arms on either the A (arm I) or B tests (arm II); the arms acted as blind controls for each other. In the classical trial the complete intervention on B tests (arm II) was compared with a control arm without any intervention on B tests (arm III).

Results

The trials with the block design yielded statistically significant changes in the numbers of A tests ordered ($p=0.013$), but not in the numbers of B tests ordered ($p=0.29$). In the classical design, the complete intervention reached a marginal significant change in the B tests ($p=0.068$). The Hawthorne effect was the same for both arms of the block design. In the classical design, the effect could to some extent be attributed to the Hawthorne effect.

Conclusion

Our block design had a surplus value compared with the classical design, in that it allowed us to control for the Hawthorne effect. Suitable use of block designs may further our knowledge of non-specific effects in quality improvement research.

Key words

Quality research; design; randomised controlled trial; Hawthorne; non-specific effects

Acknowledgements

The authors gratefully acknowledge the financial contribution to the study by the Dutch Health Care Insurance Council.

Introduction

To bridge the gap between evidence-based medicine and practice, we need to learn more about effective quality improvement interventions for implementation of research findings in daily practice.[1-6] Evaluating these interventions demands for rigorous methodology and is both complex and challenging.[7-13] Randomised controlled trials are considered as the most robust method of assessing such strategies, because randomisation normally ensures that known and unknown biases are distributed evenly between the trial arms.[4,8] When evaluating interventions aimed at improving clinical practice, however, there are a number of non-specific effects which may influence estimations of the effect of an intervention in randomised trials. These include positive attention effects, caused by participants knowing that they are the subject of a study, but also negative, de-motivating effects caused by being allocated to a control rather than to an intervention group. These non-specific effects are currently grouped together under the name 'Hawthorne effect'. If these are imbalanced across study groups in a quality improvement trial, the resulting effect estimates may be biased [14-20] However, there is relatively little empirical data about the potential influence of such non-specific effects.

Randomised controlled trials utilising balanced incomplete block designs should balance such non-specific effects.[2,20,21] The simplest such design is a 2 x 2 balanced incomplete block design in which subjects are randomised into two groups. Group 1 receives the intervention for condition A and provides control data for condition B, whereas group 2 receives the intervention for condition B and provides control data for condition A. The design is balanced because it ensures that all participants receive the same intensity of intervention and data collection and should therefore balance any non-specific effects. The design is incomplete because not all participants receive the complete intervention for both conditions.[16,17]

We just finished a trial evaluating an intervention aimed at improving GPs' test ordering performance. Since the Hawthorne effect may influence the outcome of this trial, the present paper determines the possible added value of block designs compared with classical designs in controlling for the Hawthorne effect. Therefore, the results of a simple classical two-arm trial are compared with the results of a 2 x 2 balanced incomplete block design within the same study. To our knowledge, it is one of the first empirical studies in the health care setting that tries to determine whether balanced incomplete block designs provide different results compared to simple two arm trials.

Methods

Background

The aim of the study was to evaluate the effect of a new intervention to improve general practitioners' test ordering. The intervention involved: personalised, comparative feedback; dissemination of and education on national, evidence-based guidelines; and regular quality improvement meetings in small, existing local GP teams. Two groups of targeted tests were identified including tests for cardiovascular,

upper abdominal and lower abdominal problems (A tests) and tests for pulmonary, non-organ related and degenerative joint complaints (B tests). More details have been provided elsewhere. [22]

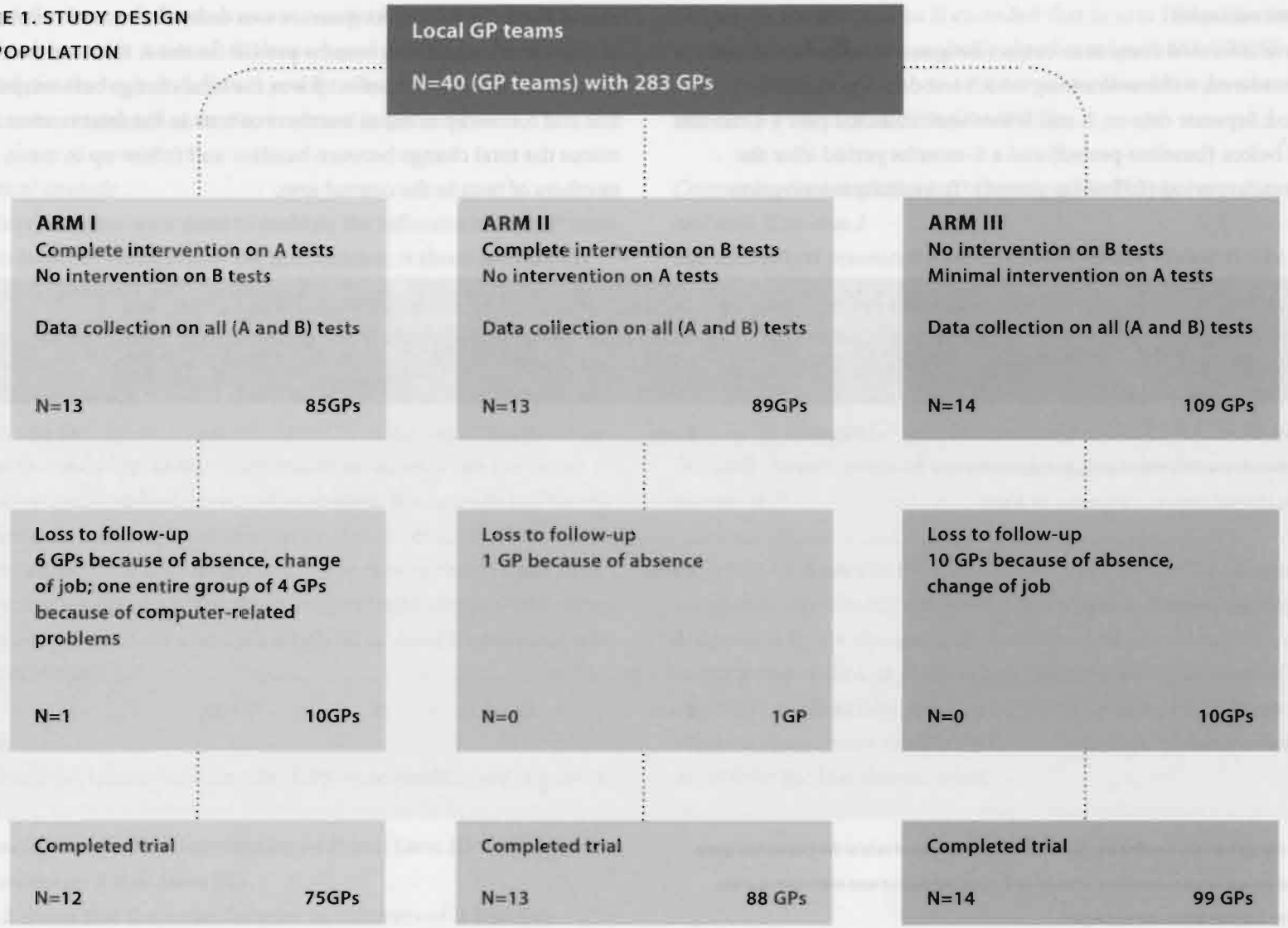
Evaluating the design

The trial was a three-arm cluster randomised trial with the local GP team as the unit of randomisation. **Figure 1** shows the design of the study. GP teams randomised to arm I received the intervention for A tests, while GP teams in arm II received the intervention for B tests. GP teams in arm III received a minimal intervention for A tests. Data on A and B tests were collected from all arms of the trial. Arms I and II represented a 2 x 2 balanced incomplete design, while arms II and III represented a simple two- arm randomised trial of the intervention on B tests. Consequently, our design was a combination of a classical two-arm RCT and a RCT with a block design. **Table 1** shows the hypotheses of the different trials and the possible value concerning the Hawthorne effect.

TABLE 1 DESIGNS, HYPOTHESES PER TRIAL AND POSSIBILITY TO ENSURE EQUAL HAWTHORNE EFFECT ACROSS THE ARMS

TRIAL COMPARING	DESIGN	HYPOTHESIS	CONTROL HAWTHORNE
ARM II- III Complete intervention for B tests. No intervention on B tests (control). Minimal intervention on A tests.	Classical	If the numbers of B tests in the intervention arm decrease in accordance with the guidelines and no change occurs in the control arm the intervention has a favourable effect, but without controlling for the Hawthorne effect.	no
ARM I-II Complete intervention on A tests. Using B tests as control.	Block	If the numbers of A tests decrease in accordance with the guidelines and no change in numbers of B test occurs, the intervention has a genuine effect.	yes
ARM II-I Complete intervention on B tests. Using A tests as control.	Block	If the numbers of B tests decrease in accordance with the guidelines and no change in the numbers of A test occurs, the intervention has a genuine effect.	yes

FIGURE 1. STUDY DESIGN AND POPULATION



Outcome measures.
GPs gave informed consent to extract data on the volume of A and B tests ordered, without knowing which test data were actually being collected. Separate data on A and B tests were collected over a 6-months period before (baseline period) and a 6-months period after the intervention period (follow-up period). To evaluate intervention effects, the following effect measure was defined: the total number of requested tests per six months per GP for the A tests and for the B tests. The intervention effect β was the total change between baseline and follow-up in mean numbers of tests in the intervention arm minus the total change between baseline and follow-up in mean numbers of tests in the control arm.

TABLE 2 TRIAL EFFECTS OF THE INTERVENTION FOR THE TOTAL GROUP OF TESTS BY WEIGHTED ANALYSIS OF COVARIANCE.

TRIAL COMPARING	DESIGN	EFFECT ON	β	S.E. β	P	95% C.I.
ARM II- III Complete intervention for B tests. No intervention on B tests (control).Minimal intervention on A tests.	Classical	B tests	-32	17	0.068	-66; 2.4
ARM I-II Complete intervention on A tests.	Block	A tests	-33	13	0.013	-59; -7.0
ARM II-I Complete intervention on B tests.	Block	B tests	-19	18	0.29	-55; 17

β = total change between baseline and follow-up in mean numbers of tests in the intervention group minus total change between baseline and follow-up in mean numbers of tests in the control group, corrected for baseline differences and region.

Because the overuse of tests is a common problem in general practice, and consistent with national, evidence-based guidelines for test ordering for the included clinical problems a decrease in the numbers of tests was considered as an improvement in patient care.

Statistical analysis

Multilevel analyses were done to evaluate the influence of the GP team level in terms of the effects of the intervention. A three-level model was used with the GP team as level 3, GPs as level 2, and the assessment of the numbers of tests as level 1. For reasons of power, effects were analysed using analysis of covariance with the follow-up assessment of the numbers of tests as dependent variable and the baseline numbers of tests and the region, which was found to be an important determinant of test ordering, as the independent variables. Since the point estimation and standard deviation were about the same in multilevel analyses as in the analysis of covariance at individual GP level, no correction for GP teams was needed. Inspection of the residual plots showed that weighted analysis was necessary in the classical trial design, and, to be consistent, we also used weighted analyses for the trials with the block design.

Results

Forty local GP teams, including 283 GPs, were randomised (**Figure 1**).

Comparing the complete intervention on B tests (arm II) with no intervention on B tests (arm III)

Table 2 shows that the mean decrease in numbers of B tests per GP per six months in arm II exceeded that in arm III by 32 tests, a difference which was marginally significant (p= 0.068, 95% CI –66; 2.4). No changes in numbers of A tests were found between arm II and III (β = -3, p= 0.80, 95% CI –26; 20).

Comparing the complete intervention in a block design: arm I vs. arm II and arm II vs. arm I

The trial effects presented a differentiated picture (**Table 2**). The effect on A tests was that the decrease in arm I (complete intervention on A tests) exceeded that in arm II (the control arm) by 33 tests per GP per six months (95% CI –59; -7). The effect on the B tests in arm II (complete intervention on B tests) exceeded that in arm I (the control arm) by 19 tests per GP per six months (p= 0.29, 95% CI –55; 17). Detailed clinical results of the block design trial have been reported elsewhere. [22]

The effect on B tests in the classical trial was found to be larger and marginally significant, compared to the effect on B tests in the block design trial. In the classical trial the effect of the complete intervention on the group B clinical problems was larger than in the trial with the block design. The classical two-arm trial on B tests overestimated the effect compared with the trial in the block design, because it did not control for the Hawthorne effect.

Discussion

The most important conclusion is that our pragmatic design, a combination of a classical and a block design, evaluating an intervention for improving GPs' test ordering performance, was able to evaluate a complex, multifaceted intervention in detail. Our design allowed us to conduct three two-armed randomised trials, one classical and two block design trials. Both arms of the block design trial involved the same intervention, and all aspects of data collection was given, so the Hawthorne effect was equal in both arms. The larger effect in the classical trial was probably due to the Hawthorne effect and not to the intervention itself: the more attention given, the greater the effect. However, in research trying to evaluate such a quality intervention, which includes attention as an important element, it appeared to be difficult to assess the magnitude of the Hawthorne effect, because it is hardly possible to differentiate between the amount of attention given as part of the intervention and the Hawthorne effect. Hence, because it seemed possible to control for the Hawthorne effect, the block design proved to have an added value compared with the classical design.

The present study had some limitations. Financial and organisational restrictions prevented us from including a real control arm, with no intervention at all. Such a design might provide more accurate answers to the problem of the magnitude of the non-specific effects. Moreover, we did not handle other non-specific effects, such as the fact that contact between intervention and control physicians can influence outcome, as physicians talk about the quality strategy under study, an effect commonly known as 'across subject contamination effect' or 'leaking effect'.[8,23] We presume this effect is not that large, because GPs normally do not discuss test-ordering performance amongst themselves and the teams were located in different regions. For the same reason, we also assumed that GPs were really blind for the fact that we collected more test ordering data than the data they were intervened on, and for the fact that their colleagues received another intervention or the same intervention on other clinical problems. Finally, although we accounted for baseline measurements in the analysis of covariance we did not address the 'ceiling effect': the fact that there is little room for improvement in high performance scores.[20] The ceiling effect in this study may have been important, since GPs in the Netherlands order considerably fewer tests than GPs in other countries.[24]

We can conclude that in evaluating an intervention to improve or change performance, it seems important –where possible- to use a block design. Since this means that the GPs in both arms of the intervention are subject to the same level of intervention, the Hawthorne effect is equalised across the two arms. It is necessary to monitor carefully how the blocks of a block design are composed, as there must not be any interference between the two blocks. Contamination, another non-specific effect, may be a major threat to the validity of block designs, which may occur when participating physicians improve their performance not only for topics under study, but also for related ones. Therefore, it is necessary to gather more data than where the intervention is focused on to be able to control for this effect. Although a block design can result in a complex study, a major benefit of block designs is obviously the possibility to do two randomised

trials with the same intervention on the same cohort of participants. Finally, since the willingness of GPs to participate in quality improvement research may be a problem, amongst other reasons, because of the chance of being randomised to a control arm, a block design ensures that all professionals are randomised to an 'intervention' arm. We conclude that our block design proved to be an effective design to evaluate our improvement strategy, allowing us to control for the Hawthorne effect, although further studies on non-specific effects in quality research are certainly required.

References

1. Bero LA, Grilli R, Grimshaw JM, Harvey E, Oxman AD, Thomson MA. Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. The Cochrane Effective Practice and Organization of Care Review Group. *BMJ* 1998;317 :465-8.
2. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;321:694-6.
3. Mason J, Wood J, Freemantle N. Designing evaluations of interventions to change professional practice. *J Health Serv Res and Policy* 1999;4: 106-11
4. Freemantle N, Wood J, Crawford F. Evidence into practice, experimentation and quasi experimentation: are the methods up to the task? *J Epidemiol Community Health* 1998;52:75-81.
5. Grimshaw JM, Shirran L, Thomas R, Mowatt G, Fraser C, Bero L, et al. Changing provider behavior: an overview of systematic reviews of interventions. *Med Care* 2001;39::Ii2-45.
6. Grol R. Personal paper. Beliefs and evidence in changing clinical practice. *BMJ* 1997;315: 418-21.
7. Grol R, Baker R, Moss F. Quality improvement research: understanding the science of change in health care. *Qual Health Care* 2002;11:110-1.
8. Grimshaw J, Campbell M, Eccles M, Steen N. Experimental and quasi-experimental designs for evaluating guideline implementation strategies. *Fam Pract* 2000;17 :S11-6.
9. Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. The implications of adopting a cluster design are still largely being ignored. *BMJ* 1998;317:1171-2.
10. Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. *Fam Pract* 2000;17:192-6.
11. Stephenson J, Imrie J. Why do we need randomised controlled trials to assess behavioural interventions? *BMJ* 1998;316:611-3.
12. Kerry SM, Bland JM. Analysis of a trial randomised in clusters. *BMJ* 1998;316:54.
13. Mollison J, Simpson JA, Campbell MK, Grimshaw JM. Comparison of analytical methods for cluster randomised trials: an example from a primary care setting. *J Epidemiol Biostat* 2000;5:339-48.
14. Cochran WG, Cox GM. *Experimental designs.* New York: Wiley, 1957.
15. Campbell DT, Stanley J. *Experimental and quasi-experimental design for research.* Chicago: Rand McNally, 1966.
16. Cook TD, Campbell DT. *Quasi-experimentation: design and analysis issues for field settings.* Chicago: Rand McNally, 1979.
17. Murray DM. *The design and analysis of group randomised trials.* Oxford: Oxford University Press, 1998.
18. Donner A, Klar N. *Design and analysis of cluster randomization trials in health research.* London: Arnold, 2000.
19. Buck C, Donner A. The design of controlled experiments in the evaluation of non-therapeutic interventions. *J Chronic Dis* 1982;35:531-538.
20. Eccles M, Grimshaw J, Campbell M, Ramsay C. Research designs for studies evaluating the effectiveness of change and improvement strategies. *J of Qual and Saf Health Care* 2003.
21. Eccles M, Steen N, Grimshaw J, Thomas L, McNamee P, Soutter J, et al. Effect of audit and feedback, and reminder messages on primary-care radiology referrals: a randomised trial. *Lancet* 2001;357:1406-09.
22. Verstappen WH, Van Der Weijden T, Sijbrandij J, Smeele I, Hermsen J, Grimshaw J, et al. Effect of a practice-based strategy on test ordering performance of primary care physicians: a randomized trial. *JAMA* 2003;289:2407-12.
23. Winkens RA, Knottnerus JA, Kester AD, Grol RP, Pop P. Fitting a routine health-care activity into a randomized trial: an experiment possible without informed consent? *J of Clin Epid* 1997;50:435-9.
24. Leurquin P, Van Casteren V, De Maeseneer J. Use of blood tests in general practice: a collaborative study in eight European countries. Eurosentinel Study Group. *Br J Gen Pract* 1995;45:21-5.

CHAPTER IX General discussion and conclusions

Wim HJM Verstappen

Introduction

Transparency and improvement of the care provided to patients are important topics in current discussions about the future of health care services. There is much debate about the best way to improve patient care and there is a demand for new approaches that fit well within the routines of clinical professionals. One of the aspects that remain to be identified is the best method to influence the test ordering behaviour of general practitioners (GPs) or primary care physicians (PCPs).[1-7] This thesis focuses on improving test ordering performance in primary care by means of an innovative, multifaceted strategy, which was systematically developed by means of a study on the determinants of test ordering and a systematic literature review. The final strategy consisted of the following elements: transparency through personalised graphical feedback, dissemination of and group education on national, evidence-based guidelines, and small group quality improvement meetings in existing local GP groups. These GP groups are an existing part of the infrastructure of Dutch GPs, who collaborate in a specific region and share patient care outside office hours. We performed a randomised clinical trial to evaluate the effects, cost effects and feasibility of this multifaceted and innovative strategy. This chapter summarises the main findings and discusses methodological aspects of our research project. It ends with recommendations for implementing this innovative strategy on a larger scale and recommendations fur further research.

Main findings

Study on determinants of test ordering

The study on determinants of test ordering behaviour explicitly included context-related factors at GP group and regional levels. This enabled us to focus on the variation in GPs' test ordering behaviour in relation to both professional and context-related determinants, such as practice type, different ways of organising test requests or experience with feedback on test ordering data. We found large differences in test ordering between the five regions included in the analysis. Three determinants were found to be independently associated with the volume of tests, namely the GPs' involvement in developing guidelines, working in a group practice, and having had more than one year of experience of using a problem-oriented form. Nevertheless, the determinant study could explain only part of the interregional variation.

Literature review

A systematic review of intervention to improve physicians' test ordering performance, carried out for EPOC Cochrane Library, revealed the following. Although the results were heterogeneous due to differences in the type or intensity of the intervention and the setting, and because of methodological differences between studies, there were some consistent findings. Probably, different strategies are needed for modifying overuse of tests versus improving appropriateness of test ordering behaviour. It is not clear that single strategies have less impact versus multifaceted strategies, but it seems important to focus the intervention at both the professional and the context. Audit and feedback seem effective for both decreasing absolute test rate and improving

appropriateness of test use. Reminders by computer aided decision support improve the appropriateness of test use. Outreach visits, patient-mediated interventions and small group quality improvement deserve more attention. The literature considered social influence as a potentially important strategy to improve test ordering behaviour.

Effects of a strategy combining feedback, guidelines and small group quality improvement

A strategy was developed on the basis of current insights into effective change in patient care and was evaluated by a multicentre randomised controlled trial with a balanced, incomplete block design after one year. The relatively short intervention period (6-months) resulted in a substantial reduction in the total numbers of tests ordered, as well as in the numbers of inappropriate tests ordered. These reductions and the latter reduction in particular were regarded as a quality improvement in terms of test ordering, because these changes were in agreement with the recommendations in national evidence-based guidelines. The multifaceted strategy was also compared with a single strategy, namely 'classic' feedback only, to evaluate the added value of the small group quality meetings. In the arm that received the complete strategy, there was a statistically significant and clinically relevant decrease in the numbers of tests, in line with the national evidence-based guidelines, compared with the feedback only arm. The inter-doctor variation in the numbers of tests ordered decreased in both arms, but more so in the total strategy arm. Important elements of the strategy are the discussions on test ordering data, the national guidelines, the personal interaction with colleagues, and the role of the medical co-ordinator of the diagnostic centre (a special facility where GPs can order laboratory, imaging and function tests without referring patients for specialist care). Merely sending feedback reports to GPs without additional activities, such as peer discussion or other strategies that fit in well with everyday practice, seemed to have little impact.

We also developed a framework to evaluate the costs of quality improvement strategies in the absence of clinical patient data. [8-11] Running costs, development costs, and scientific costs were determined for the added value trial. The new strategy was found to result in greater cost reduction than feedback alone.

Process evaluation of quality improvement strategies is seen as a necessary addition to effectiveness studies to assess important elements of change. [12] It was therefore important to determine the extent to which the intended elements of the multifaceted strategy were accepted and actually implemented by the participants and to assess their opinion on the key elements of the feedback and the small group quality meetings. The strategy was favourably evaluated in a prospective process evaluation during the trial. Although it was found that organising the intervention required considerable effort, it did not take up much of the participating GPs' time (three 90-minutes meetings per year). All local GP groups expressed a desire for continuation of the meetings after the experiment. None of the participating GP groups regarded it as a problem to discuss individual feedback reports openly. By relating their personal feedback reports to existing national evidence-based guidelines, and by assessing barriers to and incentives for change,

GPs were able to develop individual and group plans for change to improve their test ordering performance. The three key elements of the quality meetings, mutual feedback by working in pairs, discussing national guidelines and making individual and group plans for change, were implemented at a satisfactory level in both intervention years. GPs appreciated this approach, because it was closely related to their everyday work routine.

Methodological considerations

The determinant study

The determinants study investigated the influence of context-related determinants not only at practice level but also at the level of local GP groups, including differences between GP groups in patterns of collaboration, as well as the regional level, including differences between regions in quality improvement programmes or methods of organising test requests. We performed a multilevel multivariable regression analysis on our baseline data, linked with survey data on professional characteristics and with data on context-related factors. It was relatively easy to retrieve data on laboratory tests for the baseline performance assessment from the diagnostic centres. It was more difficult to retrieve reliable imaging and function test data, since the registration of these data was not always computerised. The survey had a high response, probably because the medical co-ordinators of the diagnostic centres repeatedly encouraged GPs to fill in the questionnaire.

Although we studied context-related factors at the regional level, such as differences between regions in quality improvement programmes, we should perhaps have paid more attention to a wider set of organisational and socio-cultural determinants to allow us to explain more of the observed interregional variation. However, our finding that the GPs' involvement in developing guidelines and their experience with using the problem-oriented order form were independent predicting variables was new and valuable. Other organisational and socio-cultural determinants could include regional morbidity figures, methods of organising test requests or cultural or religious characteristics of the patient population. One region was a former mining region, which is well known for its above-average levels of cardiovascular and pulmonary diseases, but this fact alone was unlikely to explain why the mean number of tests ordered per GP was almost three times as high in this region as the region with the lowest mean number of tests ordered per GP. Additionally, local experts claimed that the regional Department of General Practice of the regional university had an important social influence on the behaviour of the GPs in the region with the lowest mean number of tests ordered by strongly advocating rational test ordering.

The effect studies

The outcome measures in the effect studies were volume data, the total number of tests ordered per GP per six months, and the number of specific tests defined as 'inappropriate' according to the guidelines. Unfortunately, we could not use clinical data, but since the evidence-based guidelines recommend a reduction in the total numbers of tests included in the trial, the observed decrease can be interpreted as a quality improvement. Moreover, there is empirical evidence that a

general reduction in test use in primary care does not lead to more referrals or substitution of care.[13] Although the guidelines recommend a reduction in total test ordering, it cannot be concluded that the numbers of individual tests should always decrease. In monitoring diabetic patients, for instance, it was necessary to increase the number of serum creatinine tests and the tests for lipid management. Furthermore, the new guidelines on COPD recommend GPs to order more lung function tests.

In general, however, the focus of our intervention was on a decrease in the volume of tests, in accordance with the national evidence-based guidelines. This means that there was a potential danger of the intervention resulting in underuse of tests. Nevertheless, we do not think underuse has been a real threat, because our strategy aimed at preventing inappropriate use, which includes both overuse and underuse of tests. GPs discussed their feedback data and related them with guidelines, and if these guidelines recommended an increase in a specific test, such as lung function tests, GPs or GP groups made plans for ordering more tests.

A decrease in 'inappropriate' tests may definitely be regarded as a quality improvement. According to the guidelines, tests were regarded as inappropriate for the associated clinical problems for various reasons, for instance because the results of these tests do not influence the treatment, because there is a high likelihood of false-positive results, because there are better alternatives, or because there are negative side-effects to the tests, such as unnecessary radiation exposure.

We also considered using 'diagnostic yield' as a kind of measure of quality of test ordering. The diagnostic yield of a test is the percentage of positive test results divided by the total numbers of this specific test ordered, which might be a valid parameter in diagnostic testing. However, this measure was found to be too difficult to interpret for the participating GPs, because in general practice there may be other reasons to order a test than diagnostic purposes. In monitoring diabetes, for instance, a high positive yield of glucose tests would indicate a poor quality of diabetes control.

A new framework for cost studies of quality of care interventions

Since only intermediate effect measures, rather than patient outcome measures, were available, a real cost-effectiveness study was not possible. Since negative effects on patient outcomes were hardly to be expected in this quality improvement strategy, our cost analyses can be seen as cost minimisation analysis.[14] However, it would be possible to measure the effects of quality improvement strategies and the cost effects at patient level, by following patients for several years in terms of clinical outcome parameters. This could extend the present study and make it a true cost-effectiveness study.

The cost minimisation analysis undertaken in our study was done from a societal perspective, but the perspective of the physicians involved may also be important. Their perspective is especially important judging the likelihood that the approach will be widely implemented.

Although our study focused on costs and cost reductions, expressed in monetary units, the new strategy may be expected to yield non-monetary benefits as well, such as improvement of the GPs' clinical knowledge and performance in other areas than test ordering or improved collaboration between GPs, benefits which we did not measure. However, it is difficult to quantify these important benefits in such a cost analysis.

Another important aspect of measuring the costs of the intervention was that of the opportunity costs for the time spent by GPs in attending the quality meetings. These costs should normally be included, because such time is not available for other activities. However, it remains debatable if these opportunity costs have to be taken into account because Dutch GPs are obliged to engage in continuous medical education programmes, such as our quality strategy, for up to 40 hours a year, and health insurers to some extent include compulsory continuous medical education in the GPs' fees.
Finally, it was found to be difficult to differentiate between development and research costs. Cost analyses usually exclude research costs, and it is debatable how development costs should be handled, because some of these costs will be necessary when implementing a strategy at a wider scale.

The value of a block design in evaluating professional performance
Evaluating professional performance demands a rigorous methodology.[15-22] Our research question made a double-blind design infeasible, because it was not possible to blind subjects in our study for the new strategy, although it may be possible to blind subjects for a routine strategy.[23,24] The main effect study used a balanced, incomplete block design.[25] This design is called balanced because both arms in the block design received the same type of intervention. It is called incomplete since the content of the intervention differed between the arms.
One of the main problems in effectiveness studies on quality improvement strategies is how to cope with non-specific effects, such as the Hawthorne effect: the mere fact that a professional is taking part in a trial and is being observed will stimulate him or her to perform better, that is, more in accordance with what is expected.[26-28] Since both arms of the block design in our study involved the same intervention and were identical in all aspects of data collection, the Hawthorne effect was assumed to be equal in both arms. Although it would have been interesting to determine the magnitude of the Hawthorne effect, we were not able to do so, because this study did not include a control group without intervention.

General discussion

Much attention is currently being invested in a systematic development of new quality improvement strategies. Facilitators and barriers have to be determined to map interventions, because we need to be cautious about strategies designed behind a desk.[29] Our determinant study and systematic review intended to provide valuable input for the design of a strategy. The determinant study showed that a problem-oriented laboratory order form had a significant impact on test ordering. This justified the use of the problem-oriented order form as an inclusion

criterion for diagnostic centres that wanted to participate. The same study found that the influence of the local GP group was small, but that many GPs mentioned the social influence of colleagues as an important determinant of clinical performance. Moreover, the literature showed that social influence, was a potentially important element of the new strategy. Social influence from respected colleagues or opinion leaders might have a greater effect on changing practice routines than traditional medical education activities.[30,31] While the social influence of the group was already incorporated in our strategy, the influence of the medical co-ordinators of the diagnostic centres probably was large. These medical co-ordinators, who provided the feedback on test ordering and could as such be regarded as experts on this topic, functioned as respected opinion leaders and stakeholders in the field of quality improvement in test ordering.[5,32,33]

The effect studies

The effect studies revealed modest, yet statistically significant changes in test ordering behaviour. The two sets of clinical problems included in the block design trial were chosen deliberately to prevent contamination between the clinical problems in the two arms. To ensure comparability, both blocks included one clinical problem with tests that are important for monitoring patients (cardiovascular topics and COPD/asthma) and two clinical problems for which tests mainly serve to exclude or confirm certain diseases. It was not possible to prevent all contamination, as, for example, the clinical problem of upper abdominal complaints and general fatigue / vague complaints are not entirely independent. We do not think, however, that this contamination biased the results. Such contamination would tend to reduce the difference between the intervention and control condition in terms of the change in the numbers of tests ordered before and after the intervention, so the actual effect may even have been underestimated.

In the block design trial, there were obvious differences in effect between the two arms: test ordering for all clinical problems in the first arm showed significant improvement, whereas test ordering for all clinical problems in the second arm tended to improve as well, although the change failed to reach statistical significance. The reasons for this intriguing difference in outcome between the two arms are not entirely clear.

One possible explanatory factor is the following. The most important clinical problem in terms of prevalence in the second arm was vague complaints / general fatigue. During the small group quality improvement meetings GPs discussed the test ordering guidelines on these problems, which recommend delayed testing in patients with vague complaints. Many GPs reported that they found it difficult to implement this guideline, and indeed, our study hardly found any change after the intervention.[34-36] Confronted with such complaints, GPs probably follow fixed routines, and use laboratory tests to win time or to negotiate with patients, who often expect or demand such tests.[34] A second factor may have been that guidelines on degenerative joint complaints recommend not to order X-rays of possibly degenerate joints, because the result of such examinations does not influence the treatment. However, GPs do not always find it easy to adhere to this guideline, again because they can use these imaging investigations to

win time.[37] In addition, this guideline does not accord with the daily practice of orthopaedic surgeons, who always order X-rays when a GP refers a patient with degenerative joint complaints. A third factor is that the guidelines for COPD/asthma were updated during the intervention period, which may have caused the lack of significant change in test ordering for this clinical problem.

We were unable to study long-term effects of the intervention, and we do not know whether the effect will persist. We do not expect that the same decrease in numbers of tests ordered will be found each year. In time, the volume of tests will probably stabilise, assuming that the practice population remains stable and there is no changing in the guidelines. Ideally, there will come a moment when GPs order a specific test entirely in accordance with the guidelines, which may then may be seen as the 'benchmark' number for that specific test; with no further change required. This may imply that future quality meetings could then focus on a new set of diagnostic tests and procedures. Cost effects will not be the same each year either, as, for instance, learning effects may mean that the strategy becomes less time-consuming and less costly, while on the other hand the effect, that is, savings from the decrease in test ordering, may also become smaller.

It can be concluded that the intervention was practice-based and expensive, and led to modest but significant changes after a relatively short intervention period, while the long-term effects are as yet unknown. Some intervention studies have achieved greater changes in test ordering, sometimes using simpler interventions such as changing the order form or using quality management interventions.[4 6 38-40] Changing the order form was found to be an effective intervention in many studies and the quality management intervention focused on specific cardiac tests. In general, these favourable interventions were aimed at a few specific tests or focused particularly on knowledge improvement, rather than performance change. We think that the effect evaluations and process evaluations we applied to our strategy showed it to be a powerful and feasible, tailor-made strategy, which fits in well with routine GP practice and routine professional development in many (Western) countries. In addition, it is linked to everyday practice work and it gives GPs the opportunity to discuss their test ordering performance with colleagues on the basis of actual performance data, making discussions less non-committal. Many test ordering problems that GPs encounter in everyday practice, such as demands for tests by patients and new guidelines, can be discussed and may be solved in an open and respectful discussion among professionals.
We also expect that other health care professionals working in teams, such as medical specialists, dentists, midwives or physiotherapists, could use this strategy to improve their test ordering behaviour. This strategy also fits well with the growing need for transparent health care using clinical data to further improve clinical practice. Of course, such an innovative strategy is not the sole solution for all aspects of quality improvement in test ordering performance, and further improvement may require additional strategies. Finally, although our method was applied to test ordering behaviour, it also seems applicable to quality improvement in other aspects of general practice, such as prescribing and referral behaviour. The most important effect

of our strategy may be that it promotes collaboration in local GP groups focusing on quality improvement. Our method may result in the creation of a team of professionals instead of a collection of individual physicians, which might be a very important 'side-effect' of the strategy.

Recommendations for implementation

Our intervention was found to be effective, challenging to the participating GPs and feasible in routine practice. Despite the ease with which we were able to recruit GPs for our project, it is certainly not always easy to motivate GPs to take part in new strategies on quality improvement. First, to many GPs quality improvement is to be synonymous with efficiency, which they feel is mostly relevant from the perspective of the health insurers. Further, many GPs interpret such new quality improvement strategies as attempts to show that they are not performing properly. In such an atmosphere the attitude of GPs will be less open and more defensive. Finally, lack of time is a commonly given reason for not taking part. GPs work under pressure and it is not always possible to make time to participate in quality improvement activities. Hence, it will take some effort to motivate GPs. Obviously, quality improvement strategies are intended to improve performance, but in order to make them easier to implement, they also need to create more job satisfaction, to be challenging and to be feasible in daily practice.[41-45] The barriers mentioned above have to be addressed, for example by using respected opinion leaders, and by reliable information campaigns that focus on the benefits for both patients and GPs, such as increased job satisfaction and better collaboration with colleagues. Although considerable cost-reductions could be achieved, the new strategy was not cheap, so financial incentives could also be important in implementing the new quality improvement strategy.

There are a number of questions concerning the actual implementation of this strategy at a larger scale, such as, who should organise the test ordering quality strategy, who should chair the quality improvement meetings, and how GPs are to be compensated in the strategy? Diagnostic centres, which already exist in a quarter of the Dutch hospitals and some large cities, and where GPs can order tests without referring patients to the hospital, seem to be important structures for implementing the new strategy, as it proved to be possible to implement this strategy for two years in five regions in the Netherlands with diagnostic centres. Diagnostic centres have access to the data, and it is their task to provide feedback to GPs about test ordering. As part of the project, we developed a software program to make it easier to produce feedback reports. This program was found to be easy to implement in the diagnostic centres. The data have to be reliable, because otherwise discussions will be negatively affected, but reliability will probably become less of an issue because diagnostic centres nowadays are completely computerised. In the future it should be possible to use not only volume data, but also clinical data on adherence to guidelines to make discussions even more profound. GP organisations, hospitals and health insurers should stimulate the setting up of regional diagnostic centres in all regions.

We recommend disseminating the national evidence-based test ordering guidelines used in the trial among the Dutch GP population. Of course, these guidelines have to be updated, for example by the Dutch College of General Practitioners.
While the framework of the quality meetings appeared to be clear and workable, regional aspects, of course, may result in different approaches. The iterative aspect requires a long-term effort: how long depends on the time it will take to cope with all clinical problems and on the moment when the GPs achieve the 'benchmark' in test ordering. Our process evaluation showed that GPs preferred to discuss a new clinical subject at each meeting. Each meeting can start with an evaluation to assess whether a GP or GP group have implemented the previously made plans for change and to initiate further improvements. It is important to plan meetings two to three times a year for a lengthy period of time, because such a number of meetings can easily be scheduled into other quality improvement activities of the GP group. For example, the monthly quality meetings on prescription could also use our innovative approach and thereby give the GP group enough experience to become acquainted with it. It should then be possible to replace two or three prescription quality meetings each year with meetings on test ordering. Additionally, the process evaluation showed that six to ten GPs seemed to be the most optimum number of GPs per group for this strategy, and it was necessary to have support from trained GPs or opinion leaders who know how to use our strategy. Finally, in the case of wider implementation, financial incentives for participating GPs may be important. In the Netherlands, GPs receive a fixed amount of money per year for attending quality meetings on prescription. As regards test ordering, we recommend that GP organisations and health insurers enter into an agreement to compensate local GP groups for participating in the test ordering quality circles, while GPs then commit themselves to achieve better quality and cost reductions in test ordering. These cost reductions have to be monitored to assess the feasibility of this agreement.

Recommendations fur further research

Remarkably, the region factor was found to be an important determinant of GPs' test ordering performance. However, we were not able to determine this factor in detail, and further studies on regional variation are warranted, including socio-cultural determinants such as regional morbidity rates or religion. A better understanding of factors that influence professional practice is necessary to achieve further scientific progress. Much remains unknown about determinants of test ordering and ways how to change it. Our study may have added another level to this research, which has not been explored before yet is an important level. Future studies should include regional and local as well professional determinants. Further studies are also necessary to evaluate differences in determinants of ordering laboratory tests, imaging tests and function tests. This could lead to different quality improvement strategies for different types of diagnostic tests. Finally, the influence of the patients on test ordering needs to be further investigated, because this influence seems to be increasingly important.

Our strategy could possibly be improved by combining it with other strategies: better (i.e., problem-oriented) order forms, the use of GPs' computerised systems, and the use of (electronic) reminders. Financial incentives and other organisational interventions also seem applicable in the GP setting, and deserve more attention, as does the delegation of GPs' tasks to other professionals, such as GP practice nurses, who can order glucose tests in the context of their diabetes surveillance. Further, as mentioned above, it seems useful to study the effects of introducing our strategy in other domains of general practice, such as referral or prescription performance, and among other teams of collaborating professionals.

Although the block design can be applied to several fields of quality of care research, we do not recommend too rigorous designs in this research area. Other research methodologies should be developed for situations where such rigorous designs are not possible. Time series analyses with enough measuring points would seem a useful design for, for instance, most organisational interventions.[36]
Finally, we need valid methods to determine the heterogeneity of intervention effects and the generalisability of study results in the quality of care domain. More standardisation of intervention descriptions, outcome measures and data analysis are needed to allow fair comparisons between studies.

References

1. Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician performance. A systematic review of the effect of continuing medical education strategies. *JAMA* 1995;274(9):700-5.
2. Thomson O'Brien MA, Oxman AD, Davis DA, Haynes RB, Freemantle N, Harvey EL. Audit and feedback: effects on professional practice and health care outcomes. *Cochrane Library* 1997.
3. Thomson O'Brien MA, Oxman AD, Davis DA, Haynes RB, Freemantle N, Harvey EL. Audit and feedback versus alternative strategies: effects on professional practice and health care outcomes. *Cochrane Library* 1997.
4. Solomon DH, Hashimoto H, Daltroy L, Liang MH. Techniques to improve physicians' use of diagnostic tests. A new conceptual framework. *JAMA* 1998;280:2020-2027.
5. Thomson O'Brien MA, Oxman AD, Haynes RB, Davis DA, Freemantle N, Harvey EL. Local opinion leaders: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev* 2000(2):Cd000125.
6. Van der Weijden T, Grol R, Winkens R, Buntinx F, ter Riet G, Klazinga N. Interventions aimed at influencing the use of diagnostic tests. The relevance of attention for contextual factors.[Protocol]. *Cochrane Library* 2001.
7. Grol R, Grimshaw J. From best evidence to best practice: effective implementation of change in patients' care. *Lancet* 2003;362(9391):1225-30.
8. Mason J, Eccles M, Freemantle N, Drummond M. A framework for incorporating cost-effectiveness in evidence-based clinical practice guidelines. *Health Policy* 1999;47(1):37-52.
9. Eccles M, Mason J, Freemantle N. Developing valid cost effectiveness guidelines: a methodological report from the north of England evidence based guideline development project. *Quality in Health Ccare* 2000;9(2):127-32.
10. Mason J, Freemantle N, Nazareth I, Eccles M, Haines A, Drummond M. When is it cost-effective to change the behavior of health professionals? *JAMA* 2001;286(23):2988-92.
11. Brown CA, Belfield CR, Field SJ. Cost effectiveness of continuing professional development in health care: a critical review of the evidence. *BMJ* 2002;324(7338):652-5.
12. Hulscher ME, Laurant MG, Grol RP. Process evaluation on quality improvement interventions. *Qual Saf Health Care* 2003;12(1):40-6.
13. Winkens RA, Grol RP, Beusmans GH, Kester AD, Knottnerus JA, Pop P. Does a reduction in general practitioners' use of diagnostic tests lead to more hospital referrals? *Brit J of Gen Pract* 1995;45(395):289-92.
14. Buxton MJ, Drummond MF, Van Hout BA, Prince RL, Sheldon TA, Szucs T, et al. Modelling in economic evaluation: an unavoidable fact of life. *Health economics* 1997;6(3):217-27.
15. Stephenson J, Imrie J. Why do we need randomised controlled trials to assess behavioural interventions? *BMJ* 1998;316(7131):611-3.
16. Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. The implications of adopting a cluster design are still largely being ignored. *BMJ* 1998;317(7167):1171-2.
17. Kerry SM, Bland JM. Analysis of a trial randomised in clusters. *BMJ* 1998;316(7124):54.
18. Mollison J, Simpson JA, Campbell MK, Grimshaw JM. Comparison of analytical methods for cluster randomised trials: an example from a primary care setting. *J Epidemiol Biostat* 2000;5(6):339-48.
19. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;321(7262):694-6.
20. Grimshaw J, Campbell M, Eccles M, Steen N. Experimental and quasi-experimental designs for evaluating guideline implementation strategies. *Fam Pract* 2000;17 Suppl 1:S11-6.
21. Grol R, Baker R, Moss F. Quality improvement research: understanding the science of change in health care. *Qual Health Care* 2002;11(2):110-1.
22. Flottorp S, Oxman AD, Havelsrud K, Treweek S, Herrin J. Cluster randomised controlled trial of tailored interventions to improve the management of urinary tract infections in women and sore throat. *BMJ* 2002;325(7360):367.
23. Winkens RA, Knottnerus JA, Kester AD, Grol RP, Pop P. Fitting a routine health-care activity into a randomized trial: an experiment possible without informed consent? *Journal of Clin Epid* 1997;50(4):435-9.
24. Freemantle N, Wood J, Crawford F. Evidence into practice, experimentation and quasi experimentation: are the methods up to the task? *J Epidemiol Community Health* 1998;52(2):75-81.

25. Cochran WG, Cox GM. *Experimental designs.* New York: Wiley, 1957.
26. Roethlisberger FJ, Dickson WJ. *Management and the worker.* Cambridge: Harvard University Press, 1939.
27. Parsons HM. What Happened at Hawthorne? New evidence suggests the Hawthorne effect resulted from operant reinforcement contingencies. *Science* 1974;183:922-932.
28. Adair JG, Sharpe D, Huynh C-L. Hawthorne Control Procedures in Educational Experiments: A Reconsideration of Their Use and Effectiveness. *Review of Educational Research* 1989;59(2):215-228.
29. van Bokhoven MA, Kok G, van der Weijden T. Designing a quality improvement intervention: a systematic approach. *Qual Saf Health Care* 2003;12(3):215-20.
30. Greer A.L. The state of the art versus the state of the science. The diffusion of new medical technologgies into practice. *Int J Technol Assess Health Care* 1988;4:5-26.
31. Mittman B, Tonesk X, Jacobson P. Implementing clinical practice guidelines: social influence strategies and practitioner behavior change. *Qual Rev Bull* 1992;18:413-422.
32. Lomas J, Enkin M, Anderson GM, Hannah WJ, Vayda E, Singer J. Opinion leaders vs audit and feedback to implement practice guidelines. Delivery after previous ceasarean section. *JAMA* 1991;265(17):2202-7.
33. Borbas C, Morris N, McLaughlin B, Asinger R, Gobel F, Lomas J, et al. The role of clinical opinion leaders in guideline implementation and quality improvement. *Chest* 2000;118(2 Suppl):24s-32s.
34. Jung HP, Wensing M, Grol R. What makes a good general practitioner: do patients and doctors have different views? *Br J Gen Pract* 1997;47(425):805-9.
35. van der Weijden T, van Bokhoven MA, Dinant GJ, van Hasselt CM, Grol RP. Understanding laboratory testing in diagnostic uncertainty: a qualitative study in general practice. *Br J Gen Pract* 2002;52(485):974-80.
36. Van der Weijden T, Van Velsen M, Van Hasselt CM, Dinant G, Grol R. Unexplained complaints in general practice: prevalence and the general practioners performance. *Med Dec Making* 2002.
37. Little P, Cantrell T, Roberts L, Chapman J, Langridge J, Pickering R. Why do GPs perform investigations?: The medical and social agendas in arranging back X-rays. *Fam Pract* 1998;15(3):264-5.
38. van Walraven C, Goel V, Chan B. Effect of population-based interventions on laboratory utilization: a time-series analysis. *JAMA* 1998;280(23):2028-33.
39. Isouard G. A quality management intervention to improve clinical laboratory use in acute myocardial infarction. *Medical Journal of Australia* 1999;170(1):11-4.
40. Kendrick D, Fielding K, Bentley E, Kerslake R, Miller P, Pringle M. Radiography of the lumbar spine in primary care patients with low back pain: randomised controlled trial. *BMJ* 2001;322(7283):400-5.
41. Shortell SM, Bennett CL, Byck GR. Assessing the impact of continuous quality improvement on clinical practice: what it will take to accelerate progress. *Milbank Q* 1998;76(4):593-624, 510.
42. Shortell SM, Zazzali JL, Burns LR, Alexander JA, Gillies RR, Budetti PP, et al. Implementing evidence-based medicine: the role of market pressures, compensation incentives, and culture in physician organizations. *Med Care* 2001;39(7 Suppl 1):I62-78.
43. Tausch BD, Harter MC. Perceived effectiveness of diagnostic and therapeutic guidelines in primary care quality circles. *Int J Qual Health Care* 2001;13(3):239-46.
44. Grimshaw JM, Eccles MP, Walker AE, Thomas RE. Changing physicians' behavior: what works and thoughts on getting more things to work. *J Contin Educ Health Prof* 2002;22(4):237-43.
45. Szecsenyi J, Beyer M, Gerlach F, al e. The development of quality circles/peer review groups as a method of quality improvement in Europe.Results of a survey in 26 European countries. *Fam. Pract* 2003;20:443-52.

CHAPTER X Summary

CHAPTER I introduces the subject of this thesis, the test ordering behaviour of general practitioners (GPs). The numbers of tests ordered by GPs are increasing and many of these tests appear unnecessary according to established evidence-based guidelines. Furthermore, inter-doctor variation seems to be large. This thesis describes the variation in test ordering behaviour in primary care, provides a systematic literature review on the strategies used by others to influence physicians' test ordering behaviour, and discusses the effects and costs of an innovative strategy we developed to improve GPs' test ordering behaviour. The strategy was systematically developed on the basis of the findings of the literature review. The multifaceted strategy had an iterative character and included the following elements: personalised graphical feedback, guideline dissemination and continuous small group quality improvement meetings. An important feature of the graphical feedback reports was a comparison between the behaviour of individual GPs and that of their colleagues. Mutual feedback by working in pairs, discussing guidelines, and drawing up plans for change were important features of the small group quality improvement meetings.
The meetings were organised in local GP groups. Local GP groups are an existing part of the infrastructure of Dutch GPs working together in a specific region. One of their tasks is to organise care during out-of-office hours, while continuing medical education is another important activity in many of these local groups. Co-ordination of the feedback and supervision of the group meetings was provided by a diagnostic centre, a facility where GPs can order laboratory, imaging and function tests without referring patients for specialist care. One of the tasks of the medical coordinator of the diagnostic centre is to give feedback to GPs on their test ordering behaviour.

The multifaceted, innovative approach was implemented in five regions catered for by five diagnostic centres, all working with a problem-oriented test ordering form. It was evaluated in 40 local GP groups by means of a multicentre trial with randomisation at local GP group level.

CHAPTER II describes a survey study of the variation in the test ordering behaviour of the GPs that participated in the test ordering trial, which tried to establish professional-related and context-related determinants of GPs' inclination to order tests, by means of a cross-sectional analysis. The baseline data of the trial, which involved 19 laboratory and 8 imaging tests, combined in a sum score per GP per year, were analysed to assess determinants of inter-doctor variation. In a multivariable multilevel regression analysis, these data were linked with survey data on professional characteristics such as knowledge about and attitude towards test ordering, and with data on context-related factors such as practice type or experience with feedback on test ordering data. The response to the survey was 97 %. Test ordering data were available for 229 GPs in 40 local GP groups from five regions. We found that the total median number of tests per GP per year was 998 (interquartile range 663 to 1500), with large differences between the regions ($p<0.001$). Factors significantly associated with smaller number of tests ordered were, at professional level, 'individual involvement in developing guidelines' (yes versus no) and at context level 'group practice' (versus two-person or single-handed practices) and

'more than one year of experience working with a problem-oriented laboratory order form' (yes versus no). GPs who met these three criteria ordered 27%, 18%, and 41% fewer tests, respectively, than their colleagues. We concluded that, in addition to professional-oriented determinants, context-related factors are strongly associated with the numbers of tests ordered. Further studies on GPs' test ordering behaviour should include local and regional factors.

Chapter III reports on a systematic literature review of strategies to influence test ordering behaviour, applying rigorous Cochrane Collaboration methods. It was hypothesised that changing the absolute rate of test use (which in most cases meant reducing the general overuse of diagnostic tests) and improving the appropriateness of test use (usually by means of explicit guidelines for certain disease-defined patient categories) are different issues that need different strategies. The second hypothesis was, that multi-faceted strategies would generally have a greater impact than single strategies. Finally, it was hypothesised that studies evaluating strategies involving context-oriented interventions would have a greater impact than exclusively professional-oriented interventions. A total of 98 studies with 118 comparison groups were included. Overall results were heterogeneous, due to differences in the type or intensity of the intervention or the setting, or due to methodological differences between studies. Modifying the overuse of tests and improving the appropriateness of test ordering behaviour may require different strategies. In addition to professional-oriented interventions, it seems important to consider the use of interventions that focus on organisational factors. It is not clear whether single strategies have less impact than multifaceted strategies, but it seems important to focus the intervention on both the professional and the context. Audit and feedback seem effective in decreasing absolute test rates as well as in improving the appropriateness of test use. Reminders by computer aided decision support were found to improve the appropriateness of test use, while outreach visits, patient-mediated interventions and small group quality improvement deserve more attention.

Chapter IV evaluates the strategy of combining feedback, guideline dissemination and small group quality improvement on the basis of a multicentre randomised controlled trial with a balanced, incomplete block design. The primary outcome measure was the total number of tests ordered for three different clinical problems per GP per six months. Arm I consisted of 13 groups receiving the strategy on three clinical problems, viz., cardiovascular diseases, upper abdominal complaints and lower abdominal complaints, while arm II consisted of 14 groups that received the same strategy, but concentrating on three other clinical problems, viz., COPD / asthma, general malaise / vague complaints and degenerative joint complaints (see chapter I, figure 1). The ordering volume of all tests related to the six clinical problems was monitored in both arms. The GPs were blinded for the intervention on the three clinical problems included in the other arm. In agreement with existing national, evidence-based guidelines, decreases in the total numbers of tests ordered as well as in the numbers of tests ordered per clinical problem and for some specified inappropriate tests were regarded as quality improvements. Analysis of covariance showed that

in arm I, the decrease in the total numbers of tests relating to cardiovascular diseases, upper abdominal complaints, and lower abdominal complaints was greater than in arm II, the difference being 67 tests more per GP per six months (p =0.01). For the GPs in arm II the mean change in the numbers of tests for COPD / asthma, general malaise / vague complaints and degenerative joint complaints was greater than that in arm I, the difference being 28 tests (p=0.22). In both arms, there was a reduction in the ordering of specified inappropriate tests, although the reduction was not significant for the GPs in arm II. The new strategy, focusing on guidelines and interaction and feedback between GPs, thus seems an effective tool for improving GPs' diagnostic testing.

CHAPTER V assesses the added value of small peer group quality improvement meetings for improving test ordering behaviour compared to one of the elements of the strategy, viz., simple feedback, on its own. This research question was evaluated by comparing arm I (see chapter IV) with a third arm including GPs receiving feedback on the same three clinical problems as in arm I (cardiovascular diseases, upper abdominal complaints, and lower abdominal complaints). The complete strategy was applied in 13 GP groups with 85 GPs (arm I), while 14 GP groups with 109 GPs received feedback only (arm III) (see chapter I, figure 1). Analysis of covariance showed that in arm I the decrease in the mean total number of tests (51 fewer tests per GP per six months) was far more substantial than that in the feedback arm (p=0.0049). Five tests deemed 'inappropriate' for the clinical problem of ' upper abdominal complaints' showed a greater decrease in arm I than in the feedback arm, the difference being 13 tests per GP per six months (p=0.0015). Inter-doctor variation decreased more in arm I. This implies that if audit and feedback are to be effective, they need to be integrated in an interactive, educational environment.

CHAPTER VI provides a framework for cost evaluations of quality improvement strategies. Cost analyses were done on the trial arms reported in Chapter V, that is the complete strategy, involving feedback, guidelines, and small group quality improvement, versus feedback only. Regular costs, development costs, and scientific costs were determined. Costs per GP of the new strategy were € 92.70 per six months in the total strategy arm, € 17.10 per six months in the feedback arm. An analysis of covariance was performed with the mean costs per GP per six months after the intervention as the dependent variable, and the costs of tests at baseline and the district as independent variable. The total strategy arm achieved a mean cost reduction of € 301 per GP per six months (p=0.001), while the feedback only strategy GP saved € 161 per GP per six months. Within the proposed framework, it is imperative to account for non-monetary benefits. We concluded that our strategy is a useful quality instrument. In line with the cost analysis framework for this kind of behavioural intervention, it seems useful to implement this strategy on a larger scale

CHAPTER VII evaluates the use and applicability of a multifaceted strategy to improve GPs' test ordering behaviour by means of a prospective process evaluation during the first and second years of the trial. All 193 GPs participating in arm I and arm II received the planned six feedback reports. Data from 156 quality meetings of 26 local GP

groups showed a participation rate of 81% (95% CI: 77%-85%) in the first year and 73% (95% CI: 68%-77%) in the second year. The three key points of the quality meetings, viz., mutual feedback by working in pairs, drawing up individual plans for change and drawing up group plans for change, were performed satisfactorily in both intervention years. In the first year, GPs expressed their level of satisfaction in a score of 7.55 on a scale of 0 - 10 (95% CI 7.46-7.64); the average score in the second year was 7.51 (95% CI 7.30-7.74). We concluded that the strategy is a feasible tool for continuing improvement of GPs' test ordering behaviour, which fits in well with local and regional quality improvement efforts.

To assess the value of balanced incomplete block designs in quality improvement research, and their capacity to control for the Hawthorne effect, **Chapter VIII** evaluates the study design of our trial. Local GP groups were randomised into to three arms. The GP groups in arm I received the total strategy to improve the quality of their test ordering, focusing on tests used for cardiovascular diseases and upper and lower abdominal complaints. GP groups in arm II received an identical intervention, but focusing on tests used for COPD / asthma, general malaise / vague complaints, and degenerative joint problems. GP groups in arm III received a minimal (feedback) intervention on the same tests as the GPs allocated to arm I (chapter I, figure 1). The numbers of tests related to all six clinical problems were monitored in all arms. The GPs were blinded for the interventions in the other arms. Three 2-arm comparisons were made, two within the block design, between arm I and arm II, and one with a classical design, between arm II and arm III. The block design involved analysing intervention-specific effects on changes in the number of tests ordered controlling for any Hawthorne effect. Since the GPs in both arms of the block design were subject to the same level of intervention, the Hawthorne effect was assumed to be equal in both arms. To gain insight into other potential threats to the study's validity, data on tests ordered for COPD / asthma, general malaise / vague complaints and degenerative joint problems were compared for the GPs in arm II, who had received the complete intervention with the GPs in arm III, who had only received a minimal intervention on the other three clinical problems. In the classical design the effect could to some extent be attributed to the Hawthorne effect. We concluded that the block design had a surplus value compared with the classical design. Clever use of block designs may further our understanding of non-specific effects in quality improvement research.

Chapter IX presents the general discussion and conclusions of the entire research project and the lessons to be learnt from it for the intended nation-wide implementation. Of course, innovative strategies like these are not the ultimate solution for all aspects of quality improvement in test ordering performance, and further improvement may require additional strategies. Outcome measures were volume data: the total number of tests. Unfortunately, we were unable to use clinical data. Furthermore, no long-term effects were studied. Nevertheless, the new strategy seems an innovative and practicable, efficient and cost-efficient quality instrument which can be usefully integrated within local and regional quality improvement programmes in an attempt to consistently improve GPs' test ordering behaviour.

CHAPTER XI Samenvatting

Dit proefschrift behandelt het diagnostisch aanvraaggedrag van huisartsen: het klinisch handelen van huisartsen met betrekking tot het aanvragen van aanvullend diagnostisch onderzoek, op het gebied van laboratorium, beeldvormend en functie onderzoek. In vele Westerse landen, inclusief Nederland, vragen huisartsen steeds meer diagnostische testen aan, terwijl volgens de evidence-based richtlijnen een deel daarvan overbodig is. Het proefschrift besteedt verder aandacht aan de grote interdoktervariatie tussen huisartsen wat betreft het aanvragen van diagnostische tests en een literatuur review behandelt strategieën om het aanvraaggedrag van artsen te beïnvloeden.

Hoofdstuk I geeft een globaal overzicht van dit proefschrift. Om het aanvraaggedrag van huisartsen te verbeteren, dus het realiseren dat meer aanvragen volgens bestaande richtlijnen worden aangevraagd, werd op systematische wijze een innovatieve strategie ontwikkeld, genaamd het DTO: Diagnostisch Toets Overleg. Deze meervoudige strategie bestond uit een combinatie van persoonlijke, grafische feedback rapporten, richtlijnen verspreiding en intercollegiale toetsingsbijeenkomsten. De vergelijking van het aanvraaggedrag van de individuele huisarts met zijn collega's uit de HAGRO en uit de regio was een belangrijk kenmerk van de schriftelijke feedbackrapporten.
De belangrijke kenmerken van de intercollegiale toetsingsbijeenkomsten waren de open bespreking van de feedback rapporten in tweetallen aan het begin van de bijeenkomsten, discussies over de nationale richtlijnen en het maken van individuele en groepsvoornemens op het gebied van aanvullende diagnostiek. De strategie heeft een continu karakter omdat het belangrijk is te evalueren of individuele en groepsvoornemens inderdaad leiden tot daadwerkelijke veranderingen van het aanvraaggedrag.
De meeste aandacht in dit proefschrift gaat uit naar de effecten van deze nieuwe strategie op het klinisch aanvraaggedrag van huisartsen en welke kosten en kostenbesparingen deze interventie met zich mee bracht. Deze effecten en kosteneffecten werden door middel van een gerandomiseerde studie onderzocht in een grote huisartsenpopulatie van ongeveer driehonderd huisartsen, samenwerkend in 40 huisartsengroepen (HAGRO's) in vijf regio's in Nederland. Behalve waarneming voor elkaar, scholen veel HAGRO's gemeenschappelijk na, b.v. in veel HAGRO's vindt tegenwoordig toetsing over prescriptiecijfers plaats (Farmacotherapeutisch Overleg: FTO). De coördinatie van de strategie, feedback en de supervisie, organisatie van de toetsingsgroepen lag bij de verschillende diagnostische centra. Een diagnostisch centrum is een instituut, dat meestal verbonden is aan een ziekenhuis waar huisartsen laboratorium, beeldvormend en functie onderzoek kunnen aanvragen zonder hun patiënten te verwijzen. Ongeveer een kwart van de ziekenhuizen in Nederland heeft momenteel een dergelijk diagnostisch centrum. Het geven van feedback aan adherente huisartsen over hun aanvraaggedrag is een van de taken van de medisch coördinator van het diagnostisch centrum.

Hoofdstuk II behandelt de variatie in het diagnostisch aanvraaggedrag tussen huisartsen. Professionele en contextgerelateerde determinanten die deze variatie zouden kunnen verklaren werden onderzocht. Daarvoor werd een cross-sectionele analyse gedaan van de som van 19 laboratorium en 8 beeldvormende onderzoeken, verzameld in de

vijf diagnostische centra in 1997. De samenstelling van de onderzoekspopulatie maakte analyses op een drietal niveaus mogelijk: huisarts / praktijk niveau, HAGRO-niveau en regio. In een multivariate, multilevel analyse werden deze aanvraagdata gekoppeld aan gegevens uit een enquête onder de deelnemende huisartsen over hun professionele houding ten opzichte van het aanvragen van diagnostische onderzoeken en met contextgerelateerde gegevens zoals het praktijktype of de ervaring met feedback vanuit een diagnostisch centrum. 229 Huisartsen konden in deze studie worden ingesloten. De respons op de enquête was 97%. Het totale aantal testen per huisarts per jaar was 998 (P25-P75: 663 tot 1500), met grote verschillen tussen de vijf regio's. Op professioneel niveau was 'actieve betrokkenheid bij het maken van richtlijnen' (ja/nee) en op praktijk niveau 'groepspraktijk' (vergeleken met solo- of duo-praktijken) en op regionaal niveau 'meer dan 1 jaar ervaring met het probleemgeoriënteerd laboratoriumformulier' (ja/nee), geassocieerd met respectievelijk 27%, 18% en 41% minder aanvragen. De conclusie luidt dat behalve de professionele determinanten, ook ander contextgerelateerde determinanten van invloed zijn op het aanvraaggedrag. Toekomstige studies zullen zeker rekening moeten houden met locale en regionale factoren.

Hoofdstuk III beschrijft de resultaten van een systematische literatuur review, volgens de richtlijnen van de Cochrane Collaboration, van strategieën om het aanvraaggedrag van (huis-)artsen te beïnvloeden. Een drietal hypothesen werden onderzocht. Ten eerste: of het veranderen van het absolute aantal testen, meestal door overdiagnostiek te verminderen, andere strategieën zou vergen dan pogingen om het juist gebruik van diagnostische testen (meestal door expliciete richtlijnen) te bewerkstelligen. Ten tweede is het de vraag of meervoudige strategieën mogelijk meer effect zouden hebben dan enkelvoudige en als laatste of interventies die zich richtten op de context een meerwaarde zouden hebben vergeleken met interventies die zich alleen richtten op de professional. Achtennegentig studies met 118 vergelijkingsgroepen werden geincludeerd.
De resultaten waren niet eenduidig, omdat de interventies niet steeds vergelijkbaar waren en er veel methodologische verschillen bestonden tussen de studies. Behalve algemeen geaccepteerde regels zoals het zorgen dat de te onderzoeken strategie aansluit bij de praktijk en met name rekening houdt met de weerstand tegen verandering die bij professionals vaak bestaat, is ook het doel van de interventie van belang: vermindering van het overmatig diagnostisch handelen vs. meer aanvraaggedrag volgens de richtlijnen. Deze laatste twee doelen vergen inderdaad verschillende strategieën. Enkelvoudige strategieën bleken niet minder impact hebben dan meervoudige.
Het lijkt wel belangrijk om te focussen op zowel de professional als de context. Intercollegiale toetsing en feedback lijken zowel het aantal overbodige testen als de 'kwaliteit van het diagnostisch handelen' positief te beïnvloeden. Computerondersteunende reminders verbeteren de kwaliteit van het aanvraaggedrag.
Het blijkt steeds belangrijker ook belangrijk organisatorische interventies te onderzoeken. Verder lijken strategieën die gebruik maken van patiëntenoordelen en intercollegiale toetsing zeker meer aandacht behoeven in wetenschappelijk onderzoek omdat ze in potentie een positieve invloed hebben.

In **Hoofdstuk IV** wordt de DTO-strategie geëvalueerd in een multicentre gerandomiseerd experiment met een gebalanceerd, incomplete blockdesign. De primaire uitkomstmaat was het aantal testen dat een huisarts aanvraagt per half jaar voor drie verschillende klinische beelden. De studiepopulatie bestond uit 26 HAGRO's. Arm I bestond uit 13 groepen die de totale interventie (feedback, richtlijnen en intercollegiale toetsing) kregen over testen behorende bij de drie klinische beelden cardiovasculaire ziekten, bovenbuikklachten en onderbuikklachten. Arm II bestond uit 14 HAGRO's die dezelfde interventie ondergingen met betrekking tot drie andere klinische beelden COPD/astma, vage klachten en degeneratieve gewrichtsafwijkingen. (Zie hoofdstuk I, figuur 1) Van alle deelnemende huisartsen werden de aantallen testen van alle zes klinische beelden gemonitored.

De huisartsen uit de ene arm waren blind voor het feit dat de andere groep dezelfde interventie onderging maar met betrekking tot drie andere klinische beelden. Volgens de bestaande nationale, evidence-based richtlijnen werd een daling van het totaal aantal testen waarop geïntervenieerd werd en een daling per klinisch beeld opgevat als een verbetering van de kwaliteit van het aanvraaggedrag. Verder werden enkele 'overbodige' testen gedefinieerd die volgens de richtlijnen niet meer door huisartsen hoefden te worden aangevraagd. Covariantie analyses lieten zien dat voor huisartsen in arm I de daling in aantallen testen voor de klinische beelden cardiovasculaire ziekten, bovenbuikklachten en onderbuikklachten per huisarts per half jaar gemiddeld 67 meer was dan voor huisartsen in arm II (p=0.01). Van huisartsen in arm II daalde het aantal testen voor de klinische beelden COPD/astma, vage klachten en degeneratieve gewrichtsafwijkingen met 22 meer dan voor huisartsen in arm I maar die verandering was niet significant (p=0.22). Ook de overbodige testen daalden in beide armen, hoewel die daling voor de huisartsen in arm II niet significant was. De conclusie was dat de nieuwe strategie, die zich richtte op het gebruik van richtlijnen en sociale interactie en feedback tussen huisartsen, een effectief kwaliteitsinstrument kan zijn om het aanvraaggedrag van huisartsen te verbeteren.

In **Hoofdstuk V** wordt de meerwaarde onderzocht van de richtlijnen en intercollegiale toetsing op het gebied van de verbetering van het diagnostisch aanvraaggedrag, vergeleken met klassieke feedback. Het design was een multicenter trial met randomisatie op HAGRO-niveau. De totale strategie werd in 13 groepen met 85 huisartsen uitgevoerd (arm I, dezelfde als uit hoofdstuk IV), terwijl de feedback strategie in 14 groepen met 109 HAGRO's werd gedaan (arm III). Deze huisartsen kregen feedback over drie dezelfde klinische beelden (cardiovasculaire ziekten, bovenbuikklachten en onderbuikklachten), waarover de huisartsen in arm I de DTO-strategie ondergingen (Zie hoofdstuk I, figuur 1). Volgens de richtlijnen kon een absolute daling van het aantal testen opgevat worden als kwaliteitsverbetering.
Covariantie analyses lieten een significante daling van gemiddeld 51 testen per huisarts zien vergeleken met de feedback arm (arm III). Vijf 'overbodige' testen voor het klinisch beeld bovenbuikklachten gaven een significante gemiddelde daling van 13 testen meer per huisarts per half jaar dan bij de huisartsen die alleen feedback kregen. De inter-doktervariatie daalde meer in de arm die de totale interventie kreeg

dan in de feedback arm.Vergeleken met het alleen maar toezenden van feedbackrapporten, verbeterde de DTO-strategie het aanvraaggedrag van huisartsen duidelijk meer en meer consistent. Dat betekent dat toetsing en feedback effectiever zijn als ze geïntegreerd zijn in een interactieve en educatieve omgeving.

In **Hoofdstuk VI** wordt een raamwerk gegeven voor kostenevaluaties van kwaliteitsbevorderende strategieën. Een kostenevaluatie werd gedaan met dezelfde trialarmen als in hoofdstuk V. Lopende kosten, ontwikkelingskosten en researchkosten werden vastgesteld. Per huisarts in de totale interventie arm (arm I) kostte de nieuwe strategie € 92.70 per half jaar, en in de feedback arm (arm III) waren de kosten voor de feedback strategie € 17.10 per huisarts per half jaar. Covariantie analyses met de gemiddelde kostenreductie per huisarts als onafhankelijke variabele en de gemiddelde kosten per huisarts bij de nulmeting en de regio als onafhankelijk variabele, gaven een significante hogere kostenreductie van € 144 per huisarts per half jaar voor de totale interventie arm vergeleken met de feedback arm (p=0.001). In het raamwerk behoren ook niet-geldelijke voordelen meegenomen te worden. De conclusie was dat de DTO-strategie een waardevol kwaliteitsinstrument is en dat bovendien de kosten en kostenreducties van deze nieuwe strategie het wenselijk maken verder te onderzoeken hoe deze op grotere schaal te implementeren.

In **Hoofdstuk VII** wordt het gebruik en de toepasbaarheid van de DTO-strategie beschreven met behulp van een prospectieve procesevaluatie tijdens de interventieperiode van twee jaar.

Alle 193 huisartsen van arm I en II kregen de geplande 1158 feedbackrapporten. Gegevens van 156 toetsingsbijeenkomsten gaven een opkomstpercentage van 81 % in het eerste jaar (95% BI: 77%-85%) en 73% (95%BI: 68%-77%) in het tweede jaar. De drie belangrijkste elementen van de toetsingsbijeenkomsten: paarsgewijze bespreking van de rapporten, relateren van het aanvraaggedrag aan de richtlijnen en het maken van individuele en groepsvoornemens werden in beide jaren voldoende uitgevoerd. In het eerste jaar gaven de huisartsen de totale strategie een 7.55 (95% BI: 7.46-7.64) op een 10-puntsschaal en 7.51 (95% BI: 7.30-7.74). Het DTO bleek implementabel in de dagelijkse praktijk en lijkt goed inpasbaar in locale en regionale nascholings- en toetsingsprogramma's.

In **Hoofdstuk VIII** wordt de meerwaarde van het blok design bepaald ten opzichte van een klassieke design met twee armen, o.a. in het omgaan met non-specifieke effecten zoals het Hawthorne-effect. Deze studie werd gebaseerd op de totale 3-armige studie. HAGRO's uit arm I kregen de totale strategie over testen, behorend bij de klinische beelden hart- vaatziekten, boven- en onderbuikklachten. HAGRO's uit arm II kregen de complete interventie over testen, behorend bij de klinische beelden COPD/asthma, algemene malaise en moeheid en degeneratieve gewrichtsafwijkingen.
HAGRO's uit arm III kregen de minimale feedback interventie over testen behorend bij de klinische beelden hart- vaatziekten, boven- en onderbuikklachten (Zie hoofdstuk I, figuur 1). In alle armen werden van alle huisartsen alle testen behorend bij een van de zes klinische beelden geregistreerd. Huisartsen waren blind voor de interventie in

de andere armen. De huisartsengroepen uit de eerste twee armen waren controlearm van elkaar. Drie 2-armige vergelijkingen waren mogelijk, twee binnen het blokdesign en een klassiek design tussen arm II en arm III w.b. de testen behorend bij de klinische beelden COPD/asthma, algemene malaise en moeheid en degeneratieve gewrichtsafwijkingen.
In het blokdesign werden interventie-specifieke effecten geanalyseerd, waarbij gecontroleerd werd voor het Hawthorne-effect. Omdat de huisartsen uit de armen van het blokdesign dezelfde mate van interventie ondergingen maar wel over verschillende klinische beelden werd het Hawthorne-effect gelijkelijk verdeeld over deze twee armen. In het blok design trad geen verbetering op voor de testen waarop niet geinterveniëerd werd. In het klassieke design kon het effect voor een deel worden toegeschreven aan het Hawthorne-effect en dus had het blok design een duidelijke meerwaarde ten opzichte van het klassieke design. Een juist gebruik van het blok design in kwaliteitsonderzoek kan de kennis van de invloed van non-specifieke effecten in kwaliteitsonderzoek verbeteren.

Hoofstuk IX tenslotte gaat over de algemene conclusies van het hele project. Conclusies uit de determinantenstudie, het literatuur review en de klinische en kosteneffecten van de DTO-strategie worden nogmaals kritisch beschouwd. Het is belangrijk te constateren dat in dit onderzoek de lange termijn effecten niet onderzocht konden worden en dat vooral kwantitatieve uitkomstmaten gebruikt werden. Het bleek (nog) niet mogelijk om klinische data te gebruiken. Natuurlijk is de DTO- strategie niet de ultieme oplossing om het aanvraaggedrag van huisartsen blijvend te verbeteren. Waarschijnlijk zijn ook andere strategieën mogelijk en nodig. Toch wordt geconcludeerd dat het ontwikkeld kwaliteitssysteem het diagnostisch aanvraaggedrag van huisartsen op een effectieve, kostenefficiënte en, in de dagelijkse praktijk toepasbare manier, kan verbeteren. Deze strategie kan zeker ook gebruikt worden voor andere vormen van intercollegiale toetsing bijvoorbeeld over verwijscijfers of prescriptiecijfers. Niet alleen huisartsen kunnen er hun voordeel doen mee doen, ook andere samenwerkende professionals kunnen (elementen uit) deze methode toepassen. Een bredere invoering van het DTO wordt aanbevolen.

Willy Dubois-Arbouw †

Tijdens mijn promotietraject was het plotselinge overlijden van Willy een bijzonder verdrietige gebeurtenis. Het is nauwelijks te bevatten dat iemand die actief is binnen het onderzoek plotseling wegvalt. De eerste vier jaar van het project was Willy onmisbaar voor het DTO-onderzoek.
Willy vertrok op 7 juni 2002 met haar man naar Italie voor een welverdiende vakantie. Die vrijdagochtend vroeg werden Willy en haar man, nog maar 10 minuten onderweg, getroffen door het noodlot. Ze kregen een auto-ongeluk waarbij Willy om het leven kwam en Theo zeer ernstig gewond raakte.

Vrolijk, vriendelijk, attent en met veel inzet deed ze vanaf begin 1998 haar werk als researchassistente. Ze toonde een grote betrokkenheid met de onderzoekers en haar collega-assistentes. Ze was een vraagbaak voor iedereen.
Willy had veel interesses en wilde zich breed ontwikkelen in haar vak.
Ze had ambities en wilde groeien als researchassistente.
Enthousiast vertelde Willy over allerlei andere zaken die haar bezighielden. Hoorn spelen in de harmonie van Vilt, de liefde voor haar hond, de tuin en Italiaans leren, want Italië was een passie van Willy en Theo samen.
Zonder haar had dit project niet kunnen slagen en ik had graag met haar dit onderzoek afgemaakt.

Willy werd slechts 38 jaar.

Dankwoord

Naar het schrijven van dit stuk heb ik jaren uitgekeken. Het dankwoord is het meest en best gelezen deel van een proefschrift. Het is inderdaad een feest om te promoveren. Promoveren doe je niet alleen, gelukkig maar, anders was ik er nooit aan begonnen. Ik wilde altijd al promoveren. Ik vond dat ik als huisarts van teveel dingen te weinig afwist en wilde me een aantal jaren bezighouden met één onderwerp. Toevallig (toeval bestaat niet?) werd het dit onderwerp. En via dit onderwerp heb ik veel geleerd over wetenschappelijk onderzoek, schrijven, publiceren maar ook over huisartsgeneeskunde en de huisartsen. Op het huisartseninstituut heb ik veel gemotiveerde jonge basisartsen en gezondheidswetenschappers gezien die onderzoek deden binnen de huisartsgeneeskunde. Prima, maar als we met z'n allen wetenschappelijke vooruitgang belangrijk vinden, moeten we ook vanuit de dagelijkse huisartsenpraktijk ervaren huisartsen stimuleren om en de kans geven te promoveren. Er is meer minder positiefs te vertellen over het instituut promoveren. Ik zou een grondige discussie hierover toejuichen.

De ongeveer 300 deelnemende huisartsen wil ik als eerste hartelijk danken. Jullie hebben geheel vrijwillig twee jaar deelgenomen aan deze studie. Jullie enthousiaste reacties gaven mij aan, nog meer dan de uiteindelijke positieve klinische effecten, dat we op de goede weg zaten. Huisartsen lopen echt voorop als we het we het hebben over het verantwoording afleggen voor het klinisch handelen.
Ik heb met bijzonder veel plezier samengewerkt met mijn (co-) promotores: Richard Grol, Jeremy Grimshaw en Trudy van der Weijden. Richard, je hebt dit dankwoord nog niet eerder gelezen. Ik weet zeker dat je er graag nog commentaar op had willen geven. Je hebt gezien dat ook het wetenschappelijk deel me gelukt is. Je was verbaasd dat mensen je streng vonden. Nou, laat ik je uit de droom helpen: je bent echt streng maar van jou kan ik het hebben. Vooral omdat je je verantwoordelijk opstelt, zonder verborgen agenda's. Ik vind het prettig met je verder te kunnen samenwerken. Ik wil ook in de toekomst bijdragen aan de verdere ontwikkeling van de WOK.
And you, Jeremy, I'm happy that you were willing to participate in this study. It was difficult to plan your visit to Maastricht today, but of course I postponed my defence so you could attend it. I will always remember our nice days in Dublin, Maastricht and Utrecht. It was an honour discussing my papers with you while you had just woke up in Ottawa, drinking your first cup of coffee. I hope to meet you at many scientific occasions.
En jij, Trudy, dankzij jou is het me echt gelukt. Nooit te beroerd om me weer vooruit te helpen; ik heb veel van je geleerd. Je moest me regelmatig afremmen en me duidelijk maken dat ik met een wetenschappelijk onderzoek bezig was en niet met het schrijven van een krantenartikel. Ik hoop nog vaak met je te kunnen samenwerken en let op je sleutels.

Ik dank mijn promotiecommissie voor de tijd die ze hebben gestoken in het beoordelen van dit proefschrift. Prof. De Leeuw, Prof. Engelshoven, Prof. Voorn, Prof. Buntinx en Prof. Koes.
Onze interventie werd gedaan vanuit vijf medisch coördinerende centra. De toetsgroepen in de verschillende regio's werden begeleid

door mijn collega-medisch coördinatoren van de diagnostische centra. Luuk van Paridon (Ede-Wageningen, Louis Reichert (Sittard), Jan Hermsen (Nijmegen), Ivo Smeele, Hans Vlek en Wim van Geldrop (Helmond). Het was plezierig om met deze enthousiaste groep samen te werken. Ik ben vele malen bij jullie op je centra geweest, het was iedere keer hartelijk en inspirerend. Jullie secretariaten hebben veel werk verricht en jullie hebben ervoor gezorgd dat we steeds responspercentages van boven de 90% hadden. Ik wil jullie daarvoor hartelijk danken. Ivo en Jan bovendien dank voor jullie bijdrage in de projectgroep en het meeschrijven van artikelen.

In die projectgroep zaten ook nog Frits van Merode, Gerben ter Riet, Marianne Meulepas en Ron Winkens. Frits, vooral in het begin hebben we veel samengewerkt rondom de kostenaspecten van onze interventie. Jouw inbreng was daarbij onmisbaar. Ik vond vooral je adequate en snelle reacties op mijn vragen en voorstellen prettig. Gerben, tijdens het onderzoek, ging je werken in Engeland en Zwitserland, en uiteindelijk naar Amsterdam. Dat je desondanks betrokken bleef bij mijn onderzoek zegt genoeg. Ik heb veel van je geleerd en vond de discussies met je diepgaand en zinvol. We hebben nog een paar klussen af te maken, daar verheug ik me op. Vooral in het begin was ook Marianne Meulepas betrokken. Zeker rond de theoretische onderbouwing van de interventie en de expertmeetings daarover. Ron, dank voor je kritische ondersteuning bij de uitwerking van de gegevens en het schrijven van artikelen.
Voor het schrijven van Engelstalige artikelen heb ik veel steun gehad van Jan Klerkx. Bedankt. En de onovertroffen voorkant van het proefschrift en de verschillende hoofdstukken is van Jaime van Eijkelenborg. Minstens duizend mensen hebben de afgelopen jaren meegewerkt aan het verwerken en analyseren van de gigantische hoeveelheid gegevens. Ik noem alleen de belangrijkste: Willy Dubois (†), Paula Vilters, Paula Rinkens, Anuschka Weekers, Jildou Sijbrandij, Arnold Kester, Frans Tan. De mensen van de verschillende afdelingen die mij aan gegevens hielpen: Ad Hoeks van het Sint Joseph Ziekenhuis, van het SCDC Helmond Helen Bilik, Cecile Smeets-Goevaers en Bea Heesakkers, van Meetpunt Kwaliteit van de DHV-Eindhoven, vooral Hennie van Bavel. Een hoogtepunt was toch het feit dat enkele medewerkers van het laboratorium uit het ziekenhuis van Sittard twee weekenden lang handmatig de aantallen labaanvragen uit 1997 van de huisartsen daar hebben geturfd. Ik blijf het ongelooflijk vinden.

Ik had jaren een onmogelijke agenda, nog steeds trouwens. Marjo van Ham en Peggy Veugen zorgden ervoor dat mijn agenda overzichtelijk bleef, tenminste voor mezelf. Ook het bestuur van het Medisch Integratie Centrum Kempenland in het toenmalig St. Joseph Ziekenhuis dank ik voor het vertrouwen dat jullie in mij stelden en de mogelijkheid die ik van jullie kreeg om dit onderzoek uit te voeren in de adherentie van het ziekenhuis.

We hadden een heerlijke kamer op onze vakgroep: Rogier Hopstaken (jij bent de volgende), Ben van Steenkiste (Ben, hoe moet dat ook alweer met Endnote?), Sjoerd Hobma (zullen we samen lunchen?) en Sandra Kuiper (jou gun ik deze kamer). Met jullie heb ik veel humorvolle momenten gemaakt. In de eenzaamheid van een promotietraject

was een uur lachen met jullie vaak een ontlading en ontspanning. Daarna kon ik er weer weken tegen. Ik had ook prettig contact met nog veel meer mensen van de vakgroep en ik vind het vervelend jullie achter te laten in een voor jullie zo onzekere periode, maar het komt echt goed. Ine Siegelaer, Jos op 't Root, Karin Vaessen, Marie-Louise Dumont, Bernadette Zinsen, Paddy Hinssen, Jelle Stoffers, Paul Zwietering, Jim Tatipata, Marga van der Aa, Piet Portegijs, Job Metsemakers, Geert-Jan Dinant, Paul Houben, Tanja Maas, Saskia Mol, Paul Knipschild, Loes van Bokhoven en alle anderen.

Er zijn buiten het onderzoek nog een heleboel mensen die me op hun manier gesteund hebben. Mijn collegae van de Commissie Wetenschappelijk Onderzoek van het NHG. Regelmatig hebben we de afgelopen jaren de vorderingen van mijn onderzoek besproken. En steeds even inspirerend als kritisch. Ook nu zie ik weer belangrijk huisartsgeneeskundig onderzoek (ontstaan) binnen de CWO waar ik graag meer van wil horen. Ook de mensen van mijn 'oude' huisartsengroep. Ik hoop dat we nog lang regelmatig bij elkaar blijven komen (tot en met onze rollatorfase?)
Mijn maatjes van de supervisiegroep: Pim, Vincent, Joost (niet meer de enige doctor), Els, WimB (nog steeds honderdmaal dank dat je me bij deze groep haalde), Albert, Jasper en Marian. Deze tent heb ik mooi alleen opgezet. En natuurlijk Toos Willemsen, onze niet-overtroffen supervisor: bedankt voor je vele wijze lessen die ik ook in mijn onderzoek goed heb kunnen gebruiken.

Rond mijn promotie mis ik mijn ouders. Jullie zouden reuzentrots geweest zijn op me. Ik vind het nu vooral jammer niet meer te kunnen zeggen hoe trots ik op jullie ben dat ik dit allemaal mede door jullie kan meemaken. Familie en vrienden: binnenkort heb ik weer tijd (?). Geert, jij vertegenwoordigt mijn vijf broers. Henk, René, we kunnen eindelijk naar Berlijn.

En veruit het belangrijkste: thuis. Thuis was er vooral veel warmte en gezelligheid. Het was altijd plezierig thuiskomen in een liefdevolle en enthousiaste omgeving. Lieve Marlie, bedankt dat je er bent en hoe je er bent, ik was (ben?) niet altijd even gemakkelijk. En Josephine, Barbara en Pieter, lieverds: de feestkleren zijn gekocht....

Het feest kan beginnen.

Wim Verstappen was born on June 7th, *1955* in Heythuysen, the Netherlands. He completed his secondary education at the Bisschoppelijk College in Roermond in *1972*. In *1973* he started his medical training at the University of Nijmegen. He graduated in *1981*, and subsequently trained as a general practitioner at the practice of P. Lichter (†) in Vierlingsbeek.
From *1985* until *1997* he worked as a general practitioner in the health centre Kersenboogerd in Hoorn. Since *1988* until *1995* he was also employed as a GP trainer by to the Department of General Practice of the Vrije Universiteit Amsterdam.
From *1993* until *1997* he also worked as a research coordinator for the Institute for Research in Extramural Medicine (EMGO) of the Vrije Universiteit.
From *1988* until *1992* he was involved in continuing professional development (CPD) as a member of the regional working group on CPD (WDH West-Friesland). From *1992* until now he is a member of the Conunittee for Scientific Research in Primary Care of the Dutch College of General Practitioners (NHG). From *1998* to *2000* he was a member of the working group preparing the NHG-guideline on Gout.
From *1977* until *2001* he worked as a medial coordinator in the St. Joseph Hospital in Veldhoven, the Netherlands. It was in this capacity that he executed the research described in this thesis. From *2001* until now he works in the same capacity for the Center of Diagnostics and Consultation in the St Jans Gasthuis in Weert.

He lives happily together with Marlie Konickx and their three kids: Josephine, Barbara and Pieter.

TOWARDS OPTIMAL TEST ORDERING IN PRIMARY CARE

To bridge the gap between evidence-based medicine and practice, we need to learn more about factors and interventions that are important for the implementation of research findings in clinical practice. There is much debate about the best way to improve patient care and there is a demand for new approaches that fit well within the routines of clinical professionals. Transparency and improvement of the care provided to patients are important topics in current discussions about the future of health care services. In many countries, the number of diagnostic tests ordered by general practitioners is growing, and inter-doctor variation is shown to be large, while according to established evidence-based guidelines, many of these tests are seen as unnecessary. This thesis describes variation in test ordering behaviour in primary care, strategies used by others to influence physicians' test ordering behaviour in a systematic literature review, and effects and costs of an innovative strategy to improve general practitioners' test ordering behaviour. The aim of the study reported in this thesis was the systematic development and assessment of an innovative and multifaceted strategy to improve general practitioners' test ordering behaviour. The multifaceted strategy had an iterative character and included the following elements: personalised graphical feedback including comparative data, guideline dissemination and continuous small group quality improvement meetings.

The new strategy seems an acceptable and feasible quality instrument to reduce the general practitioners' test ordering volume in an efficient way, and can be integrated within local and regional quality improvement programmes.