The efficacy of an automated feedback system for general practitioners

Refereed papers

The efficacy of an automated feedback system for general practitioners

Rianne Bindels PhD
Researcher

Arie Hasman PhD
Professor

Department of Medical Informatics, Maastricht University, The Netherlands

Arnold D Kester PhD
Statistician, Department of Methodology and Statistics, Maastricht University, The Netherlands

Jan L Talmon PhD
Associate Professor

Paul A de Clercq PhD
Researcher

Department of Medical Informatics, Maastricht University, The Netherlands

Ron AG Winkens PhD MD
Researcher and General Practitioner, Transmural Care Unit, University Hospital Maastricht and Department of General Practice, Maastricht University, The Netherlands

ABSTRACT

Objective An automated feedback system that produces comments about the non-adherence of general practitioners (GPs) to accepted practice guidelines for ordering diagnostic tests was developed. Before implementing the automated feedback system in daily practice, we assessed the potential effect of the system on the test ordering behaviour of GPs.

Design We used a randomised controlled trial with balanced block design.

Setting Five times six participant groups of GPs in a computer laboratory setting.

Intervention The GPs reviewed a random sample of 30 request forms they filled in earlier that year. If deemed necessary, they could make changes in the tests requested. Next, the system displayed critical comments about their non-adherence to the guidelines as apparent from the (updated) request forms.

Subjects Twenty-four randomly selected GPs participated.

Main outcome measures The number of requested diagnostic tests (17% with 95% confidence interval [CI]: 12–22%) and the fraction of tests ordered that were not in accordance with the practice guidelines (39% with 95% CI: 28–51%) decreased due to the comments of the automated feedback system. The GPs accepted 362 (50%) of the 729 reminders.

Implications Although our experiment cannot predict the size of the actual effect of the automated feedback system in daily practice, the observed effect may be seen as the maximum achievable.

Keywords: Clinical competence, clinical decision support systems, guideline adherence, practice guidelines, primary healthcare, reminder systems, test ordering

Introduction

Over the past 20 years, the number of requested diagnostic tests has increased both in primary and secondary health care and has resulted in an increased pressure on clinical laboratories.[1] Part of the increase in test consumption is understandable and can be explained by the ageing of the patient population and the growth in the number of preventive tasks in the practice of the physician. On the other hand, the increase is also caused by the demand for care from the patients and the increased availability of new diagnostic tests for which the appropriateness of use is not always clear.[2–4]

Guidelines, such as those on appropriate test ordering, are developed to improve the quality of care and reduce unnecessary diagnostic test consumption, but their implementation and use in daily practice are a problem.[5,6] Therefore it is important to develop tools that stimulate physicians' adherence to guidelines. Over the past 10 years, decision support systems have gained popularity as an implementation strategy.[7–10] These systems have the potential to change physicians' behaviour, and their test ordering in particular.[11] The advice of the system (a recommendation) describes the discrepancies between the physician's actions and the guidelines, offering recommendations for improvement.

We developed and validated an automated feedback system named GRIF (the Dutch acronym for 'Geautomatiseerde Reminders als Interactieve Feedback').[12,13] GRIF is meant to stimulate adherence to accepted practice guidelines on diagnostic tests. This system was developed to support or even replace the written feedback given by the Transmural Care Unit of the Maastricht University Hospital since 1985.[14]

The objective of the experiment described in this paper was to assess the efficacy of an automated feedback system. Efficacy is defined as the percentage of decisions made that are in line with relevant practice guidelines. To obtain a gold standard, effects should be determined in an optimal situation, without the influence of practice conditions and the demanding patient. This may show the efficacy of the system and the magnitude of effects when such an intervention is used in daily practice. Testing new automated tools by future users in a laboratory setting prior to general use in daily practice is important to prevent ineffective interventions being implemented too widely.

Methods

The GRIF system

The GRIF system consists of five parts: a knowledge base, an order entry system, a module that provides reactive support (i.e. the recommendations), a module that provides passive support and a database.[12] The knowledge base in which the recommendations are stored now contains 150 rules (recommendations) derived from accepted national and regional guidelines about various medical problems. To use the GRIF system in daily practice, the general practitioner (GP) must enter relevant medical patient data (signs, symptoms, working hypotheses, and the reason for request) and the tests to be ordered into an electronic order entry form. The medical terms have to be entered using International Classification of Primary Care (ICPC) codes.[15] A search program assists the user in the selection of the appropriate terms that are associated with an ICPC code. Due to this standardisation of terminology, our computerised rules in the knowledge base are based on ICPC codes only. Next, the reactive support module of the GRIF system reads the patient data on the electronic request form and checks whether any of the rules in the knowledge base will trigger. If a rule triggers, the corresponding recommendation is generated immediately and presented by means of a pop-up window, overlaying the interface of the electronic request form. The recommendation window contains critical comments about the requested tests as well as a link to the text of the practice guideline for more explanation (passive support module). Finally, the GP then decides to accept or reject the recommendation.

The intervention

We randomly selected 30 from the 90 GPs in the Maastricht region. Twenty-four GPs agreed to participate in a laboratory experiment. For each GP we took a random sample of request forms he/she had submitted to the transmural care unit in the preceding year. Patient information (sex, age and medical information such as working hypothesis, complaints and medical history) was entered into the electronic order entry form of the GRIF program and ICPC codes were added to the medical patient data. The request forms were anonymised so that knowledge of test results would not influence the GPs' judgements.

In a laboratory setting, the GPs were confronted with the comments of the GRIF system on their earlier ordered diagnostic tests. The intervention took place at the Maastricht University in five group sessions. To ensure that GPs still approved their own test requests, we allowed them to remove or add tests. Then the system was activated. The GP could accept or ignore the presented recommendations. The GPs had to work through the 30 cases one by one within a limited time (45 minutes) to simulate time pressure during real patient consultation. An overview of the different steps in the experiment and the items measured is presented in Figure 1.

Following this laboratory experiment, we also planned a field trial using a balanced block design. In the field trial we divided the recommendations in the knowledge base into two clusters (A and B). Half of the GPs got recommendations on cluster A, the other half on cluster B. The GPs were blinded to the remaining recommendations. Thus we could investigate the learning effect during the field trial. To make

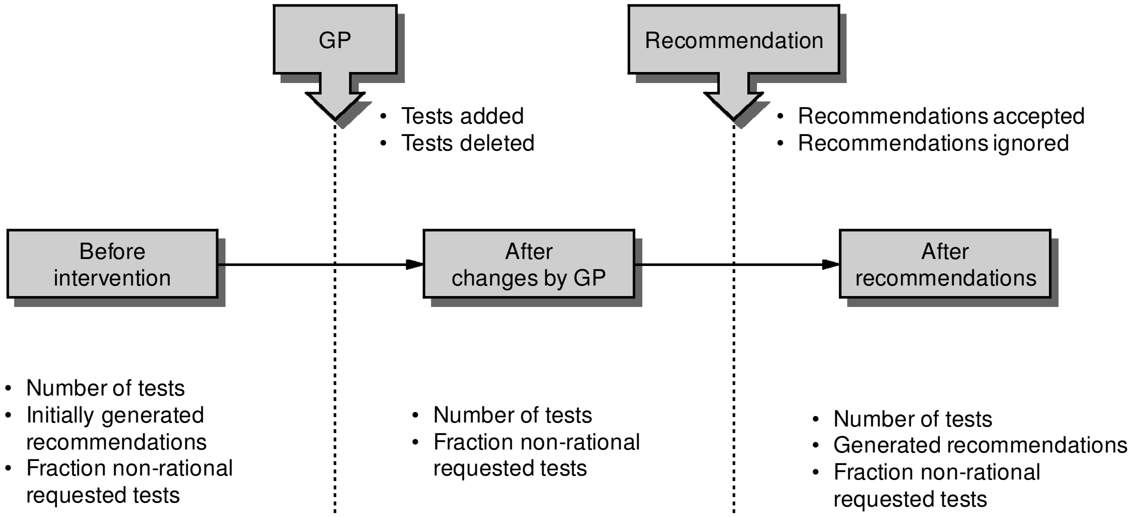
GP
Recommendation
• Tests added
• Tests deleted
• Recommendations accepted
• Recommendations ignored
Before intervention
After changes by GP
After recommendations
• Number of tests
• Initially generated recommendations
• Fraction non-rational requested tests
• Number of tests
• Fraction non-rational requested tests
• Number of tests
• Generated recommendations
• Fraction non-rational requested tests

Figure 1 Description of the different steps of the intervention

the results of the laboratory experiment comparable to the results of the field trial, we used the same design (although the blinding does not provide information in this case).

The recommendations are clustered in such a way that both clusters lead to approximately the same number of recommendations (based on results from the validation study and the fact that the severity of the diagnoses (linked to the test requests) is about equal in both clusters.[13] The diagnoses for test cluster A and B are listed in Box 1. The GPs were randomly assigned to receive either recommendations on test cluster A or on test cluster B.

Box 1 Diagnoses for clusters A and B

Cluster A

- Anaemia
- Diabetes mellitus
- Glandular fever
- Hypercholesterolaemia
- Hypertension
- Liver problems
- Urine complaints

Cluster B

- Allergy
- Diarrhoea
- Gall bladder problem
- Gout
- Infections
- Physical problems
- Rheumatism
- Sinusitis
- Thyroid gland problem

Statistical analysis

For each GP, we measured the average number of test requests at three points in time: on the original request form, after the GP's reconsiderations and after accepting/ignoring the comments of the GRIF system. We also determined for each GP the mean number of accepted and rejected recommendations per request form and studied whether tests changed by the GPs led to the disappearance of one or more recommendations. For each GP, we also calculated the percentage decrease in test numbers and the percentage decrease of tests not in accordance with the guidelines at the previously mentioned three points in time. Tests not in accordance with the guidelines were expressed as the number of recommendations not accepted divided by the total number of tests ordered. We determined for each group the absolute change in tests not in accordance with the practice guidelines: cluster A tests for GP group 1 and cluster B tests for GP group 2. First a mean value per GP was calculated, then the averages per group.

As an overall test for both clusters together, we adapted the crossover test to our model.[16] To achieve a power of 0.9, at least 22 GPs were needed to participate in this trial. The estimation in variation in request behaviour of GPs in this power calculation is based on results of a previous pilot study.[13] Since most recommendations suggest the removal of one or more tests, the direction of the effect is one-sided. Therefore we also chose to present the mean value and the 95% confidence interval (CI) of the effects. We did not determine an intracluster correlation coefficient (ICC) that takes into account that the outcomes of decisions by an individual physician may be more similar than those across study subjects.

Additionally, we studied the behaviour of the GPs in more detail and investigated correlations between changes made by the GPs and changes resulting from suggestions by the GRIF system.

Results

With respect to age, experience and sex, the 24 GPs were comparable to the whole group of GPs in our region. The mean age of the GPs was 49 years (standard deviation (SD): 5.6), the mean years of experience in general practice was 18 (SD: 7.0) and 75% of the GPs were male. In the analysis, one case of one GP was missing due to a technical problem.

Table 1 shows that the mean number of test requests decreased due to the whole intervention. The mean percentage decrease over the two GP groups was 30% (95% CI: 23% to 37%). Table 2 shows a strong decrease in the proportion of tests not in accordance with the practice guidelines (43% over the two GP groups (95% CI: 32% to 54%) for the total intervention).

The 24 GPs ordered 4196 tests (mean 5.6 tests per patient) on the original request forms. Due to reconsideration, GPs removed 545 tests (of which erythrocyte sedimentation rate (ESR) occurred 39 times, creatinine 36 times and leucocyte count 36 times) and added 80 tests (of which fasting glucose occurred 16 times and mean corpuscular volume (MCV) 13 times). On average there was a mean decrease of requested tests in the two intervention groups of 12% (95% CI: 5% to 20%).

Of the removed 545 tests, 361 tests (66%) would have generated a recommendation. Since more than one test request on one form could be related to the same recommendation, this would have resulted in a reduction of 255 recommendations. The other 184 (34%) removed tests would not have generated a recommendation. Of these, 101 should not have been removed because they were in accordance with practice guidelines. Most tests were common general tests such as glucose, creatinine, cholesterol and thyroid stimulating hormone (TSH), ordered for elderly people (age above 70) and requested in accordance with the guidelines. For 83 requested tests no guideline existed.

After the GP changed the request form, 3731 test requests remained on the request form. Due to comments of the GRIF system 457 tests were removed (leucocyte count 66 times, packed cell volume (PCV) 39 times and differential count 34 times) and 46 tests were added (fasting glucose 29 times and alanine aminotransferase (ALAT) 10 times). The comments of the GRIF system resulted in a mean decrease in the

Table 1 Mean number of test requests with corresponding standard deviation (SD) per request form per GP

	GP group 1		GP group 2	
	Cluster A: intervention	Cluster B: control	Cluster B: intervention	Cluster A: control
Before intervention	3.75 (0.90)	1.69 (0.58)	1.62 (0.44)	3.50 (1.15)
After changes by GP	3.32 (1.28)	1.49 (0.65)	1.42 (0.47)	3.09 (1.13)
After recommendations	2.67 (0.95)	1.39 (0.61)	1.09 (0.25)	3.04 (1.14)

Table 2 Mean proportion of requested tests per request form per GP that were not in accordance with the practice guidelines and corresponding standard deviations (SD)

	GP group 1		GP group 2	
	Cluster A: intervention	Cluster B: control	Cluster B: intervention	Cluster A: control
Before intervention	0.45 (0.06)	0.38 (0.06)	0.46 (0.15)	0.47 (0.09)
After changes by GP	0.41 (0.06)	0.38 (0.08)	0.44 (0.16)	0.47 (0.09)
After recommendations	0.25 (0.12)	0.38 (0.08)	0.29 (0.16)	0.47 (0.09)

number of tests ordered in the intervention groups of 17% (95% CI: 12% to 22%).

The overall difference between the intervention and the control groups was significant ($P < 0.001$).

The system generated 1420 recommendations for the 720 cases of which 729 (51%) were presented and 691 (49%) stayed hidden to the GP because of the application of the block design. Of the 729 presented recommendations, 362 (50%) were accepted and 367 (50%) were ignored. The comments of the system resulted in a mean decrease in the proportion of tests not in accordance with guidelines in the intervention groups of 39% (95% CI: 28% to 51%). For both GP groups together there was a significant decrease in the proportion of inappropriately requested tests in the intervention group compared to the control groups ($P < 0.001$).

We found four different behaviour patterns among the participating GPs. Four GPs did not change anything on the request forms in 21 or more of the 30 cases (no changes by the GP and no recommendations accepted). Nine GPs mainly changed the request forms themselves but did not follow the advice of the GRIF system. Eight GPs did not change the request form but followed the advice of the system. Finally, three GPs both changed the request form and followed the advice of the system (not necessarily in the same case).

Discussion

We have described the potential effects of an automated feedback system for test ordering by GPs. Both the number of tests ordered and the proportion of tests not in accordance with guidelines decreased.

Paper cases can be used to measure the competence of a clinician, but performance is what a physician actually does in daily practice.[17,18] Although competence scores are usually higher than performance scores, competence is a good predictor for performance.[19]

Other researchers have used simulated cases to assess the potential effect of decision support systems. They compared different forms of decision support for prescribing and for the management of breast and ovarian cancer and found they were potentially effective.[20,21] Walton *et al.* found that 35% of the recommendations of their decision support system were ignored.[21] In our study, 50% of the recommendations were ignored. A possible explanation is that GPs in the Maastricht region have received feedback on their test ordering behaviour since 1985. Their test ordering behaviour is already largely in line with the guidelines.[14]

Of the randomly selected 30 cases, some (on average two to three) cases were quite similar. This might have resulted in a learning effect during the session. Our data set was too small to detect such an effect. Nevertheless, some of the GPs indicated in group discussions held afterwards that they were more likely to accept a recommendation when it was presented repeatedly.

A disadvantage of an experiment in a laboratory setting is that the respondents may tend to give socially desired answers and act as they think the researchers want them to act. We tried to avoid this by informing the GPs that the aim of the study was to measure the user satisfaction of the system and the quality of the presented recommendations instead of focusing on measuring the performance of the GPs. Therefore, we asked them to critically review the recommendations rather than blindly accept them.

The GPs themselves deleted a relatively large percentage of the initial test requests from the request form. The majority of the tests were correctly removed (tests were not in accordance with the guidelines). Besides a second critical look at one's own test requests, the influence of the Hawthorne effect, socially desired behaviour of GPs and the absence of the patient in the laboratory setting could have played a major role in this reduction.[22]

The four behaviour patterns we found were about equally distributed over the different group sessions, indicating that they were not due to information bias. Only three GPs changed the request form themselves and followed the recommendations, a strategy that we expected for most GPs. This implies that individual GPs need different approaches to change their diagnostic test ordering behaviour.

Remarkably, we found that in cases where GPs actively changed their own request form before the system advised them, the number of tests requested not according the practice guidelines seemed to be lower than for GPs that did not actively change their request forms. It seems as if GPs that actively changed their request form know quite well which tests are appropriate or not. However, this latter finding needs to be interpreted with some caution because the analyses were performed on case level instead of on GP level.

This study clearly shows the potential of computer recommendations to achieve behaviour more in accord with guidelines in medical practice. The effects found are assumed to be most feasible.

REFERENCES

1 van Merode GG, Hasman A, Derks J, Schoenmaker B and Goldschmidt H. Advanced management facilities for clinical laboratories. *Computer Methods and Programs in Biomedicine* 1996;50(2):195–205.

2 Kassirer JP. Our stubborn quest for diagnostic certainty. *New England Journal of Medicine* 1989;320:1489–91.

3 Mandell HN. Technological imperative. Or, when your tool is a hammer, everything looks like a nail. *Postgraduate Medicine* 1983;74(2):24–6.

4 Wong ET and Lincoln TL. Ready! Fire! … Aim! *Journal of the American Medical Association* 1983;250:2510–13.

5 Woolf SH, Grol R, Hutchinson A, Eccles M and Grimshaw J. Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. *British Medical Journal* 1999;318:527–30.

6 Grol R, Eccles M, Maisonneuve H and Woolf S. Developing clinical practice guidelines: the European experience. *Disease Management and Health Outcomes* 1998;4(5):255–66.

7 Langton KB, Johnston ME, Haynes RB and Mathieu A. A critical appraisal of the literature on the effects of computer-based clinical decision support systems on clinician performance and patient outcomes. *Proceedings of the AMIA Annual Fall Symposium* 1992; 626–30.

8 Shea S, DuMouchel W and Bahamonde L. A meta-analysis of 16 randomized controlled trials to evaluate computer-based clinical reminder systems for preventive care in the ambulatory setting. *Journal of the American Medical Informatics Association* 1996;3(6):399–409.

9 Shiffman RN, Liaw Y, Brandt CA and Corb GJ. Computer-based guideline implementation systems: a systematic review of functionality and effectiveness. *Journal of the American Medical Informatics Association* 1999;6(2):104–14.

10 Sim I, Gorman P, Greenes RA *et al.* Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association* 2001;8(6):527–34.

11 van der Weijden T, Wensing M, Giffel M *et al.* Interventions aimed at influencing the use of diagnostic tests (Protocol for a Cochrane Review). In: *The Cochrane Library, Issue 1, 1999.* Oxford: Update Software, 1999.

12 Bindels R, de Clercq P, Winkens RAG and Hasman A. A test ordering system with automated reminders for primary care based on practice guidelines. *International Journal of Medical Informatics* 2000;58–59:219–33.

13 Bindels R, Winkens RAG, Pop P, Wersch JW, Talmon J and Hasman A. Validation of a knowledge based reminder system for diagnostic test ordering in general practice. *International Journal of Medical Informatics* 2001;64(2–4):341–54.

14 Winkens RAG, Pop P, Bugter-Maessen AMA *et al.* Randomised controlled trial of routine individual feedback to improve rationality and reduce numbers of test request. *Lancet* 1995;345:498–502.

15 Lamberts H and Wood M. *International Classification of Primary Care* (3e). Oxford: Oxford University Press, 1987.

16 Armitage P and Berry G. *Statistical Methods in Medical Research* (2e). Oxford: Blackwell, 1987.

17 Neufeld VR. *Assessing Clinical Competence*. New York: Springer, 1985.

18 Norman GR, Neufeld GR, Walsch A, Woodward CA and McConvey G. Measuring physicians' performances by using simulated patients. *Journal of Medical Education* 1985;60:925–34.

19 Rethans JJ, van Leeuwen Y, Drop R, van der Vleuten C and Sturmans F. Competence and performance: two different concepts in the assessment of quality of medical care. *Family Practice* 1990;7(3):168–74.

20 Emery J, Walton R, Murphy M *et al.* Computer support for interpreting family histories of breast and ovarian cancer in primary care: comparative study with simulated cases. *British Medical Journal* 2000;321(7252):28–32.

21 Walton RT, Gierl C, Yudkin P, Mistry H, Vessey MP and Fox J. Evaluation of computer support for prescribing (CAPSULE) using simulated cases. *British Medical Journal* 1997;315:791–5.

22 Deyo RA. A key medical decision maker: the patient. *British Medical Journal* 2001;323:466–7.

CONFLICTS OF INTEREST

None.

ADDRESS FOR CORRESPONDENCE

Rianne Bindels
Maastricht University
Department of Medical Informatics
PO Box 616
6200 MD Maastricht
The Netherlands
Tel: +31 43 3882295
Fax: +31 43 3884170
Email: r.bindels@mi.unimaas.nl

Accepted June 2003