Effect of a Practice-Based Strategy on Test Ordering Performance of Primary Care Physicians A Randomized Trial

Wim H. J. M. Verstappen, MD
Trudy van der Weijden, MD, PhD
Jildou Sijbrandij, MSc
Ivo Smeele, MD, PhD
Jan Hermse, MD
Jeremy Grimshaw, PhD
Richard P. T. M. Grol, PhD

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 approach.
combination of interventions.\textsuperscript{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.\textsuperscript{13,14} The aim of this strategy was to achieve sustained improvements in test ordering. Participants 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) 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).

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): 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 χ² 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). 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 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.
EFFECT OF A PRACTICE-BASED STRATEGY ON TEST ORDERING PERFORMANCE OF PRIMARY CARE PHYSICIANS

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.

**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 2. Study Population Characteristics at Individual Primary Care Physician Level**

<table>
<thead>
<tr>
<th>Individual Variables</th>
<th>Arm A</th>
<th>Arm B</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of physicians</td>
<td>85</td>
<td>89</td>
</tr>
<tr>
<td>Age, mean (SD), y</td>
<td>46.2 (6.6)</td>
<td>45.8 (5.4)</td>
</tr>
<tr>
<td>Female, No. (%)</td>
<td>14 (16)</td>
<td>15 (17)</td>
</tr>
<tr>
<td>No. of patients per physician, mean [SD]*</td>
<td>2587 (641)</td>
<td>2637 (519)</td>
</tr>
<tr>
<td>Patients &gt;65 y, % mean (SD)</td>
<td>15 (6.8)</td>
<td>13 (7.1)</td>
</tr>
<tr>
<td>Physicians with a part-time factor, % mean (SD)†</td>
<td>91 (15)</td>
<td>91 (16)</td>
</tr>
<tr>
<td>Physicians with a solo practice, No. (%)</td>
<td>43 (51)</td>
<td>48 (54)</td>
</tr>
<tr>
<td>Physicians who use computerized registration system, No. (%)</td>
<td>66 (78)</td>
<td>61 (69)</td>
</tr>
</tbody>
</table>

*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.

**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**

<table>
<thead>
<tr>
<th>Clinical Problem</th>
<th>Arm A (Intervention)</th>
<th>Arm B (Control)</th>
<th>% Change</th>
<th>% Change</th>
<th>β (SE)*</th>
<th>95% CI</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total tests</td>
<td>478 (309)</td>
<td>507 (293)</td>
<td>−12</td>
<td>0</td>
<td>−67 (19)</td>
<td>−104 to −30</td>
<td>.01</td>
</tr>
<tr>
<td>Cardiovascular/hypertension</td>
<td>293 (189)</td>
<td>290 (182)</td>
<td>−6</td>
<td>+4</td>
<td>−35 (13)</td>
<td>−61 to −10</td>
<td>.01</td>
</tr>
<tr>
<td>Upper abdominal complaints</td>
<td>165 (125)</td>
<td>192 (128)</td>
<td>−22</td>
<td>−9</td>
<td>−28 (9)</td>
<td>−45 to −10</td>
<td>.01</td>
</tr>
<tr>
<td>Lower abdominal complaints</td>
<td>20 (20)</td>
<td>25 (25)</td>
<td>−10</td>
<td>+8</td>
<td>−5 (2)</td>
<td>−9 to −1</td>
<td>.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clinical Problem</th>
<th>Arm A (Control)</th>
<th>Arm B (Intervention)</th>
<th>% Change</th>
<th>% Change</th>
<th>β (SE)*</th>
<th>95% CI</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total tests</td>
<td>640 (394)</td>
<td>724 (386)</td>
<td>−3</td>
<td>−8</td>
<td>−28 (23)</td>
<td>−74 to 14</td>
<td>.22</td>
</tr>
<tr>
<td>COPD/asthma</td>
<td>39 (31)</td>
<td>53 (27)</td>
<td>−20</td>
<td>−1</td>
<td>−1 (2)</td>
<td>−5 to 3</td>
<td>.58</td>
</tr>
<tr>
<td>General complaints</td>
<td>548 (340)</td>
<td>509 (340)</td>
<td>0</td>
<td>−5</td>
<td>−19 (21)</td>
<td>−61 to 22</td>
<td>.36</td>
</tr>
<tr>
<td>Degenerative joint complaints</td>
<td>54 (28)</td>
<td>72 (43)</td>
<td>−9</td>
<td>−19</td>
<td>−3 (4)</td>
<td>−10 to 4</td>
<td>.34</td>
</tr>
</tbody>
</table>

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.
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. 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. 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. 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. 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. 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 noncommittal. 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. 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.

Author Affiliations: Center for Quality of Care Research, Department of General Practice (Drs Verstaßen, Weijden, and Grof), Maastricht University, Maastricht, the Netherlands; Medical Integration Center Kempenland, Maxima Medical Center, Veldhoven, the Netherlands (Dr Verstaßen); Department of Methodology and Statistics, Maastricht University, (Ms Sijbrands); Center for Diagnostics and Consultation, Elkerliek Hospital, Helmond, the Netherlands (Dr Smeele); Medical Diagnostic Center of the Canisius-Wilhelmina Hospital Nijmegen (Dr Hermen); Ottawa Health Research Institute, Center of Best Practice, Institute of Population Health, University of Ottawa, Canada (Dr Grimshaw).

Author Contributions: Study concept and design: Verstaßen, van der Weijden, Smeele, Hermen, Grimshaw, Grof.

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

<table>
<thead>
<tr>
<th>Inappropriate Tests</th>
<th>Arm A (Mean) (SD)</th>
<th>Arm B (Mean) (SD)</th>
<th>β (SE)*</th>
<th>95% CI</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total tests</td>
<td>63 (75)</td>
<td>45 (41)</td>
<td>−16 (4.8)</td>
<td>−27 to −7</td>
<td>.01</td>
</tr>
<tr>
<td>BUN</td>
<td>8.7 (19)</td>
<td>7.2 (15)</td>
<td>−1 (1.3)</td>
<td>−4 to 2</td>
<td>.37</td>
</tr>
<tr>
<td>SGOT</td>
<td>7.7 (11)</td>
<td>5.5 (7.7)</td>
<td>−2 (1.4)</td>
<td>−5 to 1</td>
<td>.13</td>
</tr>
<tr>
<td>LDH</td>
<td>13 (27)</td>
<td>8.8 (16)</td>
<td>−3 (1.5)</td>
<td>−6 to −1</td>
<td>.01</td>
</tr>
<tr>
<td>Amylase</td>
<td>5.3 (13)</td>
<td>3.6 (6.9)</td>
<td>−2 (1.1)</td>
<td>−4 to −0.1</td>
<td>.04</td>
</tr>
<tr>
<td>Alkaline phosphatase</td>
<td>11 (25)</td>
<td>7.0 (11)</td>
<td>−3 (1.5)</td>
<td>−6 to −0.3</td>
<td>.03</td>
</tr>
<tr>
<td>Bilirubin</td>
<td>20 (27)</td>
<td>15 (19)</td>
<td>−6 (2.6)</td>
<td>−11 to −0.3</td>
<td>.04</td>
</tr>
</tbody>
</table>

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.

©2003 American Medical Association. All rights reserved.

(Reprinted) JAMA, May 14, 2003—Vol 289, No. 18

2411
Acquisition of data: Verstappen, Smeele, Hermsen. Analysis and interpretation of data: Verstappen, van der Weijden, Smeele, Grimshaw, Grol. Critical revision of the manuscript for important intellectual content: Verstappen, van der Weijden, Hermsen, Grol. Statistical expertise: Verstappen, van der Weijden. Obtained funding: Verstappen, van der Weijden, Grol. Administrative, technical, or material support: Verstappen, Hermsen. Study supervision: Verstappen, van der Weijden, Grol. Funding/Support: This study was supported by the Dutch Health Care Insurance Council.

REFERENCES


I know of but one freedom and that is the freedom of the mind.
—Antoine de Saint-Exupéry (1900-1944)