To test or not to test : guideline-based automated feedback on test ordering in general practice

SUMMARY

Over the past 20 years, the number of diagnostic tests that are ordered by general practitioners (GPs) has increased. A relatively large percentage of test requests in health care is ordered inappropriately when compared with practice guidelines. Although sufficient guidelines on appropriate diagnostic testing are available, it is well-known that implementation strategies are needed to make physicians to adhere to these guidelines and thus improve their test ordering behaviour. Results of previous studies are promising, but there is insufficient evidence that computerised interventions are effective implementation strategies. Apart from the scientific 'drive' to study the effects of an automated feedback system in health care, there was also a rather practical reason. Although the individual written feedback provided by the Diagnostic Centre in Maastricht improved the quality of the test ordering behaviour of GPs and this feedback method was appreciated by GPs in the Maastricht region, a more direct (related to each test order) and less laborious method of feedback was desired. Therefore, it was suggested to develop an automated feedback system that would directly assist in the test ordering process. This thesis reports on the development, validation, implementation and evaluation of an automated feedback system on test ordering in general practice.

Chapter 2 presents an overview of the literature on systems that produce real-time, patient-specific advice using on screen recommendations to physicians. This review focused on three aspects: the impact on the physician, the impact on the patient and the impact on the organisation. Besides the effectiveness of the system the physician's user acceptance of and/or user satisfaction with the system was also analysed. Thirty-eight papers were selected for complete review. These papers described different types of clinical decision support systems (CDSSs) such as reminder, alerting, suggesting or critiquing systems. Several study designs were identified but controlled before-after studies and randomised controlled trials were applied most frequently. The impact on the performance of the physician of critiquing and reminder systems in a hospital environment has been studied extensively over the past years. These studies clearly show that the physician's performance improves when these systems are used. It was concluded that the impact on the patient and the health care organisation need to be studied more extensively. In addition, researchers should focus on the question why physicians do or do not accept the advice of the CDSS.

[1] GRIF is the Dutch acronym for 'Geautomatiseerde Reminders als Interactieve Feedback' (Automated Reminders as Interactive Feedback).

In *Chapter 3* the development of the GRIF[1] automated feedback system is described. First, the selection of national or regional guidelines, the translation of the practice guidelines into computerised rules and the verification of these rules are described. The GRIF system consists of a knowledge base in which the computerised rules were implemented, as well as an order entry system and modules to provide passive and active decision support. The knowledge base consists of 149 rules applying to various medical problems. The system generates critical comments (as a result of the rules) about the appropriateness of the ordered diagnostic tests at the moment the GP orders a test that is not in line with the practice guidelines. Demographic patient data and medical data (medical history, medication and symptoms) were retrieved from the GP information system. Other relevant information had to be entered directly into the order entry system.

Chapter 4 describes the reliability of assessments of appropriateness of ordered diagnostic tests by three experts. These assessments were to be used as a reference standard in the validation process of the GRIF knowledge base (described in *Chapter 5*). A random selection of 253 request forms (with 1217 tests on it) completed by general practitioners in the Maastricht region was used. Three reviewers independently assessed the appropriateness of each requested test. Interrater kappa values ranged from 0.33 to 0.42 and kappa values of intrarater agreement ranged from 0.48 to 0.68. The joint reliability coefficient of the three reviewers was 0.66. This reliability is sufficient to review test ordering over a series of cases but is not sufficient to make reliable case-by-case assessments. In conclusion, there is substantial variation in assessments of the appropriateness of requested diagnostic tests. Computer support may be beneficial to make the process of peer review more uniform.

In *Chapter 5* the validation of the recommendations of the GRIF system is described. The comments of human experts were compared with the comments of the recommendations of the GRIF system using a retrospective random selection of 253 request forms. A panel of three expert physicians judged the requested tests independently based on their interpretations of the practice guidelines. The majority assessment of the physicians was compared with the assessment of the GRIF system. In case the system's output differed from the majority assessment the written practice guidelines were consulted.
On average 1.75 recommendations were produced per form. In total 32 (7%) of the 442 given recommendations were given incorrectly. The amount of information and the level of detail (the specificity of the terms) in which the GP describes the patient's medical status are crucial for the GRIF system to react correctly.

Chapter 6 describes the first part of the evaluation of the GRIF system. In this chapter the efficacy of the system in a laboratory setting is studied. A randomised

controlled trial (RCT) with balanced block design was used to study the potential effect of the GRIF system. The GPs reviewed a random sample of 30 request forms they had filled in and sent to the Diagnostic Center 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. Twenty-four randomly selected GPs participated.
The number of requested diagnostic tests decreased with 17% (95% CI: 12-22%) due to the comments of the GRIF system. In addition, the fraction of tests ordered not in accordance with the practice guidelines decreased with 39% (95% CI: 28-51%). The GPs accepted 362 (50%) of the 729 recommendations. Although the experiment cannot predict the size of the actual effect of the GRIF system in daily practice, this percentage of 50% is assumed to be most feasible.

The effectiveness of GRIF in daily practice is described in *Chapter 7*. Eleven GPs in two regions of the Netherlands (the regions Maastricht and Heerlen) used the system from August 2000 to July 2001. The GRIF system was implemented on the workstations at the offices of the participating GPs. The GPs were asked to use GRIF during patient consultation instead of filling in the paper request form. An analysis of usage behaviour, of the quality of information provided and of the fraction of recommendations that had been followed were presented.
During the intervention period, the GPs produced 2498 request forms using the GRIF system with 10139 tests on it. Of the 2780 recommendations, the percentage of recommendations followed varied between 3.4 and 8.3 percent depending on the type of recommendation that was given. Advice to remove a test because another - more appropriate or efficient - test was also requested and recommendations to request an alternative test were followed most frequently. Therefore, it was concluded that computerised recommendations should contain, if possible, suggestions for alternative tests to improve the application of these recommendations. The median time to generate, read and act on the feedback comments presented was 13 seconds. Entering (coded) medical patient data took GPs a relatively large part of their patient consultation time. This stresses the need for user friendly and fast data entry methods. As a result of the relatively low percentage of recommendations that were followed, creative solutions must be developed to stimulate a better adherence to these recommendations.

In *Chapter 8* the fraction of abnormal test results in tests that are ordered according to the practice guidelines are compared with the fraction of abnormal tests in tests ordered not in accordance with the practice guidelines. The outcomes of laboratory tests can seldom be equated with a diagnosis, while the outcome of imaging is often diagnostic. Hence, test requests of both domains are studied separately (the results of the study concerning imaging tests are described in *Appendix III*).

A random sample of 250 request forms that were submitted to the Diagnostic Centre by GPs was taken. These forms contained patient data (administrative and medical), the laboratory tests requested and the reason for request. The GRIF system was used to assess whether test requests are in accordance with practice guidelines.
The fraction of abnormal laboratory test results in the group of tests requested not according to the practice guidelines was significantly lower than in the group of tests requested according practice guidelines ($p=0.02$). For tests ordered to exclude a disease, the group of tests requested not according to the practice guidelines was also significantly lower than in the group of tests requested according to the practice guidelines ($p=0.04$). For tests ordered to confirm a disease there was no significant difference between the two groups ($p=0.57$).

Chapter 9 describes the user satisfaction with, the experiences with and the views on the GRIF system. This chapter combines the results from both the trial in a laboratory setting (*Chapter 6*) and the trial in daily practice (*Chapter 7*). GPs' user satisfaction was measured using a questionnaire. In addition, group discussions (in the laboratory trial) and in-depth interviews (in the field trial) were conducted to elicit the opinions about and experiences with the system. Finally, the motives for non-acceptance of the recommendations presented by the GRIF system are explored.
The results show that the GPs in the laboratory trial had more positive attitudes towards the system compared with the participants of the field trial. All discussion groups and most of the GPs in the field trial regarded receiving feedback during the test ordering process as an important advantage. The most frequently used motive not to adhere to the practice guidelines was disagreement with its the content and/or its recommendations. Apart from establishing agreement on content of the guidelines, a requirement for using GRIF in daily practice on a large scale is that more attention should be paid to the promotion, the adoption and to the stimulation of a positive attitude towards them among users.

In *Chapter 10* the conclusions of this thesis are presented and general aspects of its results discussed. This thesis shows that it is possible to generate accurate automated recommendations on tests ordering in general practice. It is advised that computerised recommendations should contain, if possible, suggestions for alternatives to improve the application of these recommendations. Furthermore, creative solutions must be developed to avoid GPs frequently ignoring the recommendations of critiquing systems. This thesis ends with recommendations for implementation in daily practice and future research.

SAMENVATTING

De afgelopen 20 jaar is het aantal door huisartsen aangevraagde diagnostische tests gestaag toegenomen. Een relatief groot percentage van deze testaanvragen binnen de gezondheidszorg worden, vergeleken met geaccepteerde richtlijnen, onterecht aangevraagd. Hoewel er voldoende richtlijnen voor het rationeel aanvragen van diagnostische tests beschikbaar zijn, is het bekend dat implementatie strategieën nodig zijn om artsen te stimuleren om zich aan deze richtlijnen te houden en dus het aanvraaggedrag van diagnostische tests te verbeteren. De resultaten van voorgaande onderzoeken zijn veelbelovend, maar er is onvoldoende wetenschappelijk bewijs dat computer-gebaseerde interventies effectieve implementatie strategieën zijn. Naast de wetenschappelijke 'drang' om de effecten van geautomatiseerde feedback systemen te onderzoeken, was er een praktische reden voor het uitvoeren van dit onderzoek. Door het Transmuraal & Diagnostisch Centrum in Maastricht (tegenwoordig Behandel & Zorgeenheid Transmurale Zorg van het Academisch Ziekenhuis Maastricht genoemd) wordt al enige jaren aan huisartsen individuele papieren feedback gegeven. Hoewel de kwaliteit van het aanvraaggedrag van de huisartsen verbeterde en deze methode gewaardeerd werd door de bij het Diagnostisch Centrum betrokken huisartsen, zocht het Diagnostisch Centrum naar een meer directe (gerelateerd aan elke testaanvraag) en minder arbeidsintensieve methode. Vandaar dat een geautomatiseerd feedback systeem, genaamd GRIF[1], ontwikkeld werd dat huisartsen tijdens het aanvraagproces kan ondersteunen in het rationeel aanvragen van aanvullende diagnostiek. Dit proefschrift beschrijft de ontwikkeling, validatie, implementatie en evaluatie van een geautomatiseerd feedback systeem betreffende het aanvragen van aanvullende diagnostiek in de huisartspraktijk.

Hoofdstuk 2 presenteert een overzicht van de literatuur met betrekking tot systemen die voor medici direct op het scherm, patiëntspecifiek advies genereren. Het hoofdstuk richt zich op drie aspecten: de invloed op de zorgverlener, de invloed op de patiënt en de invloed op de organisatie. Naast de effectiviteit van de systemen werden ook de zorgverleners' acceptatie en/of tevredenheid met het systeem geanalyseerd. Achtendertig artikelen werden opgenomen in de analyses. Deze artikelen beschreven verschillende typen beslissingsondersteunende systemen zoals reminder-, alert-, advies- of kritieksystemen. Verschillende studie designs werden geïdentificeerd, maar gecontroleerde voor-na studies en RCTs (gerandomiseerde en gecontroleerde trials) werden het meest frequent toegepast.

[1] GRIF is een afkorting voor Geautomatiseerde Reminders als Interactieve Feedback.

De invloed van kritiek- en remindersystemen op de arts in een ziekenhuisomgeving werd uitgebreid bestudeerd in de afgelopen jaren. Deze studies tonen aan dat het gebruik van dergelijke systemen een verbetering in medisch handelen teweeg brengt. Concluderend kan gesteld worden dat de invloed op de patiënt en de gezondheidsorganisatie meer frequent bestudeerd dienen te worden. Daarnaast zal meer aandacht besteed moeten worden een antwoord te vinden op de vraag waarom zorgverleners de adviezen van beslissingsondersteunende systemen al dan niet opvolgen.

In *Hoofdstuk 3* is de ontwikkeling van het GRIF systeem beschreven. De selectie van landelijke en regionale richtlijnen, de vertaling van deze richtlijnen in gecomputeriseerde regels en de verificatie van de regels wordt beschreven. Het GRIF systeem beschikt over een kennisbestand waarin regels zijn geïmplementeerd, een invoermodule en modules die passieve en actieve beslissingsondersteuning leveren. Het kennisbestand bestaat uit 149 regels die betrekking hebben op verscheidene medische problemen. Het systeem genereert kritische commentaren (als gevolg van de regels) betreffende de juistheid van aangevraagde diagnostische tests op het moment dat de huisarts een test aanvraagt die niet in overeenstemming is met de richtlijnen. Demografische patiëntgegevens en medische patiëntgegevens (voorgeschiedenis, medicatie en klachten) worden overgehaald uit het huisarts informatie systeem. Andere relevante informatie dient direct in de invoermodule ingevoerd te worden.

Hoofdstuk 4 beschrijft de betrouwbaarheid van de beoordeling betreffende de juistheid van de diagnostische tests door drie medische experts. Deze standaard zal als basis dienen bij het validatieproces van het kennisbestand van het GRIF systeem (beschreven in *Hoofdstuk 5*). Een gerandomiseerde steekproef van 253 aanvraagformulieren, die 1217 tests bevatten, werd hiervoor gebruikt. Drie beoordelaars beoordeelden, onafhankelijk van elkaar, de juistheid van elke aangevraagde test. De interbeoordelaar Kappa-waarde varieerde van 0,33 tot 0,42 en de Kappa-waarde voor intrabeoordelaar overeenstemming varieerde van 0,48 tot 0,68. De gezamenlijke betrouwbaarheidscoëfficiënt van de drie beoordelaars bedroeg 0,66. Deze waarde van de betrouwbaarheidscoëfficiënt is voldoende om de kwaliteit van het aanvraaggedrag over een serie van aanvragen te beoordelen, maar is niet voldoende hoog om de kwaliteit van afzonderlijke testaanvragen te beoordelen. Concluderend kan gesteld worden dat er een substantiële variatie bestaat in de beoordeling van correctheid van aangevraagde diagnostische tests. Computer-ondersteuning kan een toegevoegde waarde zijn om het beoordelingsproces meer uniform te maken.

In *Hoofdstuk 5* wordt de validatie van de commentaren van het GRIF systeem beschreven. Commentaren van medische experts werden in deze studie vergeleken met de commentaren van het GRIF systeem door gebruik te maken van een retrospectieve gerandomiseerde steekproef van 253 aanvraagformulieren. Een panel van drie experts beoordeelden de aangevraagde tests op basis van hun interpretatie van de richtlijnen. Het meerderheidsoordeel van de medici werd vervolgens vergeleken met het oordeel van het GRIF systeem. In die gevallen waarbij de beoordeling van het systeem verschilde van het meerderheidsoordeel van de experts, werd de tekst van de richtlijnen nogmaals aandachtig bestudeerd. Het GRIF systeem produceerde gemiddeld 1,75 commentaren per aanvraagformulier. In totaal werden 32 (7%) van de 442 gegeven commentaren niet correct gegeven. De hoeveelheid informatie op het aanvraagformulier en het niveau van detaillering (de specificiteit van de termen) waarmee de huisarts de medische gesteldheid van de patiënt beschrijft, zijn cruciaal voor het GRIF systeem om op de correcte manier te reageren.

Hoofdstuk 6 beschrijft het eerste deel van de evaluatie van het GRIF systeem. In dit hoofdstuk is de effect van het systeem in een laboratorium omgeving bestudeerd. Een RCT met een gebalanceerd blok design werd gebruikt om het potentiële effect van het systeem te onderzoeken. De huisartsen (n=24) beoordeelden een gerandomiseerde steekproef van 30 aanvraagformulieren die zij eerder dat jaar hadden ingestuurd naar het Diagnostisch Centrum. Wanneer zij het nodig achtten, konden de huisartsen de tests op het aanvraagformulier wijzigen. Vervolgens presenteerde het GRIF systeem kritische commentaren betreffende het niet opvolgen van de richtlijnen op basis van het (aangepaste) aanvraagformulier.
Het aantal aangevraagde aanvullende tests daalde met 17% (95% BI: 12-22%) als gevolg van de commentaren van het systeem. Daarnaast daalde de fractie tests die niet in overeenstemming met de richtlijnen werd aangevraagd met 39% (95% BI: 28-51%). De huisartsen volgden 362 (50%) van de 729 gepresenteerde adviezen op. Hoewel dit experiment het effect van het GRIF systeem in de praktijk niet kan voorspellen, kan geconcludeerd worden dat de geobserveerde effecten gezien kunnen worden als het maximaal bereikbare effect van het systeem in de dagelijkse praktijk.

De effectiviteit van GRIF in de dagelijkse praktijk is beschreven in *Hoofdstuk 7*. Elf huisartsen in twee regio's (de regio's Maastricht/Heuvelland en Heerlen) gebruikten het systeem van augustus 2000 tot juli 2001. Het GRIF systeem werd geïmplementeerd op de Pc's in de praktijken van de deelnemende huisartsen. De huisartsen werd gevraagd het GRIF systeem tijdens het consult met de patiënt te gebruiken in plaats van het papieren aanvraagformulier. Het gebruik van het systeem, de kwaliteit van de geleverde patiëntinformatie en de fractie commentaren die niet werden opgevolgd, werd geanalyseerd.

Gedurende de interventie periode produceerden de huisartsen 2498 aanvraagformulieren, die 10139 tests bevatten, middels het GRIF systeem. Van de 2780 commentaren die het GRIF systeem genereerde, varieerde het percentage commentaren dat werd opgevolgd tussen 3,4 en 8,3 procent afhankelijk van het type commentaar dat gegeven werd. Commentaren die adviseren een test te verwijderen omdat een andere – meer geschikte of meer efficiënte – test ook aangevraagd werd en commentaren die een alternatieve test voorstellen, werden het meest frequent opgevolgd. Dientengevolge wordt geconcludeerd dat geautomatiseerde commentaren, indien mogelijk, een alternatieve test dienen aan te bevelen om zo de toepassing van de adviezen te stimuleren.
De mediane tijd om een commentaar te genereren, te lezen en erop te reageren bedroeg 13 seconden. Het invoeren van (gecodeerde) medische patiëntinformatie kostte de huisartsen een relatief groot deel van de tijd die huisartsen hebben tijdens het patiëntconsult. Dit benadrukt de noodzaak om gebruikersvriendelijke en snelle methoden voor gegevensinvoer te ontwikkelen. Als gevolg van het relatief lage percentage commentaren dat werd opgevolgd door de huisartsen dienen creatieve oplossingen gezocht te worden om een betere overeenstemming met de richtlijnen te stimuleren.

In *Hoofdstuk 8* wordt de fractie afwijkende testuitslagen in de groep tests die aangevraagd zijn volgens richtlijnen vergeleken met de fractie afwijkende testuitslagen in de groep tests die niet volgens de richtlijnen zijn aangevraagd. De uitslag van laboratorium tests leidt zelden direct tot een diagnose, terwijl de uitslagen van beeldvormende technieken meestal diagnostisch zijn. Daarom zijn tests uit deze twee domeinen afzonderlijk bestudeerd (de resultaten van de studie betreffende beeldvormende technieken zijn beschreven in *Appendix III*). Een gerandomiseerde steekproef van 250 aanvraagformulieren, die eerder naar het Diagnostisch Centrum verzonden waren, werd genomen. Deze aanvraagformulieren bevatten (administratieve en medische) patiëntgegevens, de aangevraagde laboratorium tests en de reden van aanvraag. Het GRIF systeem werd gebruikt om te beoordelen of de aangevraagde tests in overeenstemming met de richtlijnen waren. De fractie afwijkende laboratoriumuitslagen in de groep tests die niet volgens de richtlijnen werden aangevraagd, was significant lager vergeleken met de fractie afwijkende uitslagen binnen de groep tests die aangevraagd werden volgens de richtlijnen ($p=0,02$). Voor tests die aangevraagd werden om een aandoening uit te sluiten, was de fractie afwijkende uitslagen binnen de groep tests niet volgens de richtlijnen ook significant lager vergeleken met de fractie afwijkende uitslagen van de groep tests die aangevraagd werden volgens de richtlijnen ($p=0.04$). Voor tests die aangevraagd werden om een aandoening aan te tonen was er geen significant effect tussen de twee groepen ($p=0,57$).

Hoofdstuk 9 beschrijft de tevredenheid met, de ervaring met en de mening over het GRIF systeem. Dit hoofdstuk combineert de resultaten van het onderzoek in een laboratorium omgeving (*Hoofdstuk 6*) en het onderzoek in de dagelijkse praktijk (*Hoofdstuk 7*). De gebruikerstevredenheid werd gemeten door gebruik te maken van een vragenlijst. Daarnaast werden door middel van groepsdiscussies (in de laboratorium omgeving) en diepte-interviews (in de dagelijkse praktijk) de meningen met betrekking tot en de ervaringen van huisartsen met het systeem onderzocht. Tenslotte zijn de redenen voor het niet opvolgen van de commentaren van het systeem geëxploreerd.
De resultaten laten zien dat huisartsen in een laboratoriumomgeving positiever waren ten aanzien van het GRIF systeem vergeleken met de deelnemers aan het onderzoek in de dagelijkse praktijk. Alle discussiegroepen en de meeste huisartsen die deelnamen aan het onderzoek in de praktijk, beschouwden het krijgen van feedback tijdens het aanvraagproces als een belangrijk voordeel van het systeem. De meest frequent genoemde reden om de adviezen van het GRIF systeem niet op te volgen was het oneens zijn met de inhoud van het advies in de richtlijnen. Naast het verkrijgen van overeenstemming over de inhoud van de richtlijnen, is een voorwaarde voor het gebruik van het GRIF systeem op een grotere schaal dat er meer aandacht wordt besteed aan de promotie van de richtlijnen, het volgen van de richtlijnen en het stimuleren van een positievere houding jegens richtlijnen onder gebruikers.

In *Hoofdstuk 10* worden de conclusies van deze studie gepresenteerd en worden algemene aspecten van de resultaten bediscussieerd. Dit proefschrift toont aan dat het mogelijk is om accurate, geautomatiseerde adviezen over het aanvragen van diagnostische tests in de huisartspraktijk te genereren. Geadviseerd wordt om geautomatiseerde adviezen, indien mogelijk, te voorzien van suggesties voor alternatieve tests om de toepassing van de adviezen te stimuleren. Daarnaast wordt geadviseerd om creatieve oplossingen te ontwikkelen om te voorkomen dat adviezen van kritieksystemen, zoals het GRIF systeem, veelvuldig worden genegeerd. Dit proefschrift eindigt met adviezen voor verdere implementatie van het GRIF systeem in de dagelijkse praktijk en adviezen voor verder onderzoek.