Admission and poor performance of trainees in the postgraduate GP training in the Netherlands

Margit Ilse Vermeulen

Admission and poor performance of trainees in the postgraduate GP training in the Netherlands

Thesis, Utrecht University, Utrecht, the Netherlands

ISBN/EAN: 9789039361580
Design & layout: planpuur. Utrecht
Print: Gildeprint Drukkerijen, Enschede, the Netherlands

Admission and poor performance of trainees in the postgraduate GP training in the Netherlands

Toelating en ‘poor performance’ van aios in de Nederlandse Huisartsopleiding

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
donderdag 3 juli 2014 des ochtends te 10.30 uur

door

Margit Ilse Vermeulen

geboren op 21 september 1965 te Eindhoven

Promotoren:	Prof.dr. R.A.M.J. Damoiseaux
	Prof.dr. Y. van der Graaf
Co-promotoren:	Dr. M.M. Kuyvenhoven
	Dr. H.M. Pieters

In dierbare herinnering aan mijn opa en opi.

Contents

Chapter 1

General introduction

Historical overview

Until 1973, a medical doctor could start as a general practitioner (GP) without further training immediately after graduating from medical school. In 1974, professional developments in general practice resulted in an obligatory one-year program: vocational training for general practice.[1] This training was funded by the government, which also determined the number of vacancies.[2] Admission to the training was regulated by waiting lists. Assessments during the training were carried out incidentally and implicitly; the GP trainer mainly provided a place to gain experience in daily practice and nearby functioned as a role model. Because of the job description for GPs that was presented in 1983 by the Dutch National Association of General Practitioners (Landelijke Huisarts Vereniging, LHV), further elaboration of the training seemed appropriate.[3,4] In the early 1980s, the assessment of clinical skills and communication behaviours were included in Objective Structured Clinical Examination (OSCE) settings.[5] Since 1987, a National GP Knowledge Test has been administered to all GP trainees.[6]

In 1988, the GP training program was extended to two years and admission was regulated by lottery because the waiting time had grown to several years. In 1991, a new admission procedure was introduced in which the core of the procedure was a semi-structured interview with the GP candidate conducted by a selection committee.[7] Almost all trainees who started the training became GPs, although in a few instances, a trainee was dismissed because of insufficient knowledge, skills or attitude.

In 1994, the program was elongated to three years as the evaluation of the two-year program revealed shortcomings in theoretical and practical education.[8,9] Additionally, trainees had to pass the progress qualification assessment at the end of the first year to continue with their GP training.[10] After completing the remaining two years, trainees could register as GPs. Only when serious doubts existed about a trainee's performance, the head of the training program could require remediation or dismissal of the trainee.

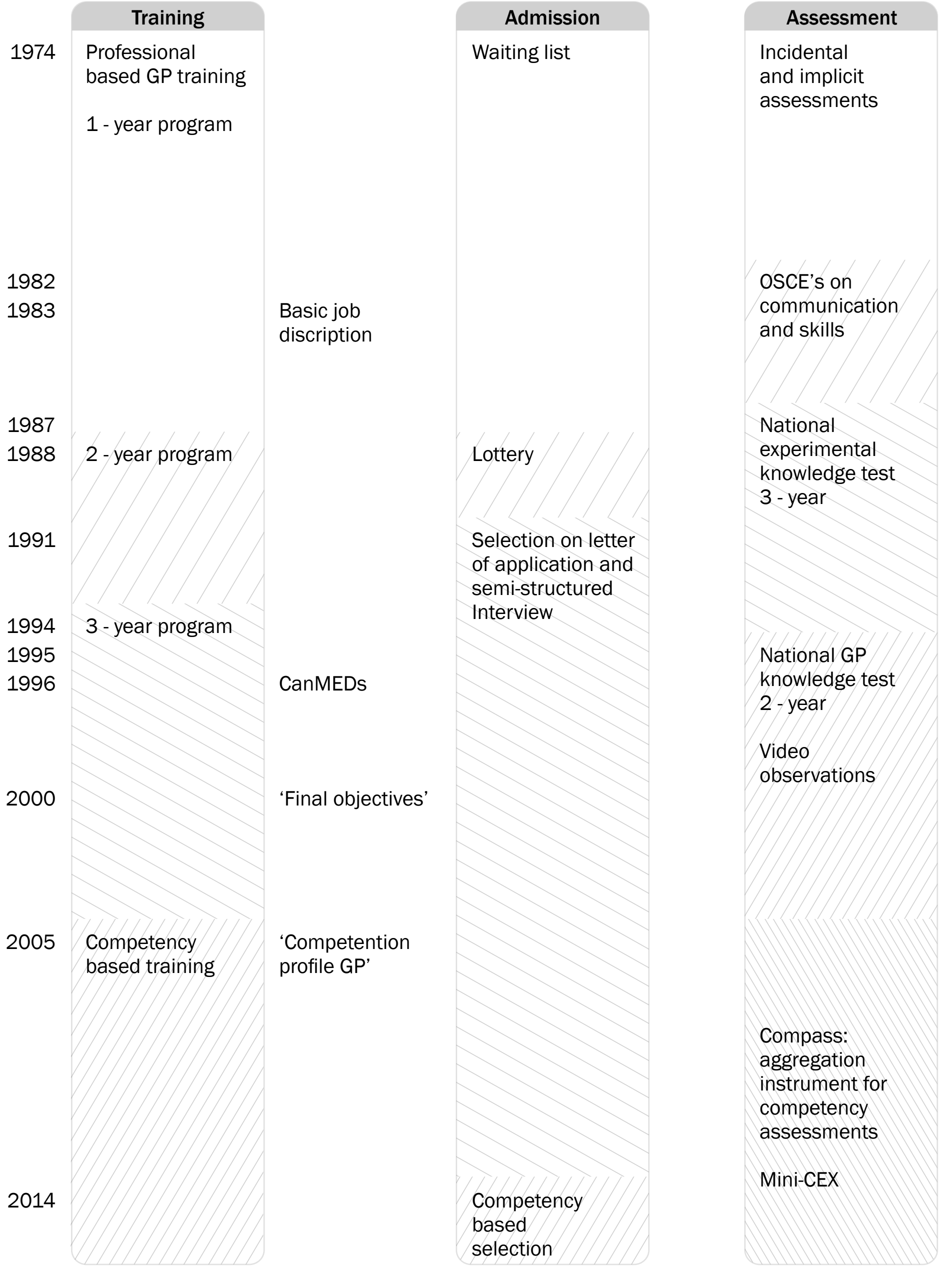
In 2005, the vocational training program was transformed into a competency based training consistent with CanMEDs, the physician competency framework.[11] The 'final objectives' and the 'competency profile of the GP' determined the content of the training.[12,13] Frequent evaluations of (sub) competencies of trainees using several assessment instruments became an essential component of the training.[14,15]

Rationale

In the context of Dutch GP training, a substantial amount of valuable research has been conducted on quality assurance and on several of the assessment instruments, such as the OSCE, knowledge tests, video observations and mini-clinical evaluation exercises.[16-24] With the inclusion of competency assessments in the assessment protocol, the formative and summative assessments have been formally regulated.[14]

However, the practice of assessment is stubborn. Aggregating all assessment information

Figure 1 Historic overview of the GP training, admission to and assessment during the GP training



	Training		Admission	Assessment
1974	Professional based GP training 1 - year program		Waiting list	Incidental and implicit assessments
1982				OSCE’s on communication and skills
1983		Basic job discription		
1987				National experimental knowledge test 3 - year
1988	2 - year program		Lottery	
1991			Selection on letter of application and semi-structured Interview	
1994	3 - year program			
1995				National GP knowledge test 2 - year
1996		CanMEDs		
				Video observations
2000		‘Final objectives’		
2005	Competency based training	‘Competention profile GP’		
				Compass: aggregation instrument for competency assessments
				Mini-CEX
2014			Competency based selection	

on a trainee, weighing the results and arriving at a valid determination regarding the trainee's competences and performance, are both difficult and troublesome. Although involuntary attrition occurs on occasion, the experience of staff members and trainers is that, in some cases, a trainee does not meet the requirements of the competence profile at the end of the training.[25,26] Do our assessment instruments and regular evaluations always guarantee the quality of the graduated trainee? Dealing with trainees, who perform poorly, is time consuming and sometimes uncomfortable. Staff members and GP trainers alike are reluctant to consider an eventual discontinuation process of a trainee because they do not want to be responsible for destroying a doctor's career.[27] At the same time, these staff members and trainers are aware of their responsibility to graduate competent GPs.

In 2005, a report on the recruitment and selection process was released with recommendations to professionalise the selection procedure and align it with the competency-based training.[28] Outside and within the medical education field, it was already known that the semi-structured interview of the selection procedure had poor predictive validity, partly because there was no congruency between the assessed personal qualities and the competencies to be acquired during the training.[29-32]

These semi-structured interviews were conducted by a staff member, a trainer and a trainee. The experience of these assessors was that the decision regarding admission was rather arbitrary and depended more on the personal judgement of the committee members, with their inherent biases, than on the competences of the candidates.[33-35] After the interviews were conducted, the interviewers met and discussed all of the candidates. Under this process, possible suitable candidates may have been rejected, and among those that were admitted, there were sometimes serious doubts about a trainee that arose during the introductory weeks of the training.

Aim of the current thesis

Based on the historical overview and rationale presented above, three aims for the current thesis were defined.

The first aim was to investigate the fairness of the selection procedure for Dutch postgraduate GP candidates. Since 1991, the selection procedure for the postgraduate GP training has been conducted within the eight training departments based on national regulations.[36] The ratio between the number of candidates and the number of vacancies differs among the eight training institutes. Thus, several questions arise. How is the admission decision reached in the eight departments? Do candidates have the same chance of being admitted by different selection committees or in different departments? Is there a difference between a candidate who applies for the first time versus a candidate who is reapplying?

The second aim was to explore poor performance of trainees in the postgraduate GP training, which often leads to stagnation and, in some cases, to the end of an individual's training program. The mean attrition rate of trainees who started the GP training program between 2002 and 2006 was 7.5%; a rate that is comparable to that of other Dutch postgraduate training programmes.[25,37] This percentage includes voluntary (5.5%) and involuntary attrition (1.9%). However, it was clear that trainees more often exhibited poor performance, and the impression was that 'professionalism' was the main role domain in which most problems occurred. Accordingly, we were interested in determining the frequency and nature of the risk factors for poor performance, as doing so would make it possible to identify poor performers early in the training.

The third aim was to develop a new standardised selection procedure for the GP training together with the postgraduate GP training of Nijmegen, as commissioned by The Dutch National postgraduate GP training (Huisartsopleiding Nederland, HON) and Foundation of postgraduate GP training (Stichting Beroeps Opleiding Huisartsen, SBOH). It was obvious that the new procedure had to be congruent with the newly adopted competency-based training. Empirical evidence and experiences from the UK were the basis of the development of the new procedure. A pilot study has been conducted to gain a first impression of reliability, validity and feasibility of the new procedure.

Outline of this thesis

Aim 1. In *Chapter 2*, we examine the selection procedure of the eight training departments, and we investigate the degree to which the department itself, the candidates' individual characteristics and the candidates' qualities explain admission to the GP training . In *Chapter 3*, we investigate the reliability of the semi-structured interviews of candidates for the Utrecht GP training , and we explore whether the results differ for candidates who apply for the first time versus those who are applying for a second or third time.

Aim 2. In *Chapter 4* and *Chapter 5*, we describe a retrospective observational cohort study of 215 trainees in the Utrecht GP training during their first year versus the entire training period, and we aim to determine how many trainees exhibit poor performance or drop out and to understand the nature of their shortcomings. In addition, we attempt to identify the risk factors for poor performance.

Aim 3. In *Chapter 6*, we describe the development process of the new competency-based selection procedure based on empirical evidence and on experiences from the UK. In *Chapter 7*, we explore and attempt to answer questions regarding the validity and reliability of the instruments of the new procedure in a pilot study alongside the current procedure.

In *Chapter 8*, we reflect on the findings of the thesis with respect to the existing literature and on current and future issues in practice and research.

The current thesis concludes with a summary of the studies in English and in Dutch.

Box: outline of postgraduate GP training

Postgraduate training in General Practice (GP training) in the Netherlands is delivered by the Departments of Family Medicine of the eight university medical centres. The program takes about 3 years, depending on relevant experience and employment rate, and has a dual character; besides the practical training, the trainee weekly attends a one day tutorial in small groups for special training and reflection, guided by staff members in the role of tutor. The content of the programme has been based on the Basic Job Description for the GP3, later by the final objectives[12] and the competency profile of the GP.[13] The first and third year comprises training in a general practice under supervision of a GP trainer. The second year is dedicated to rotations through hospital, mostly in an emergency room (six months), nursing homes (three months) and mental health institutions (three months).

References

1. Thung PJ. Raamplan 1974 van het Interfacultair Overleg der Nederlandse Faculteiten der Geneeskunde betreffende de globale doelstellingen van de artsenopleiding nieuwe stijl. Med Cont 1974;29:1017-21.
2. Capaciteitsorgaan http://www.capaciteitsorgaan.nl/
3. Springer MP, redactie LHV. Basistakenpakket van de huisarts. Bijlage bij Med Cont 1983:38.
4. Commissie Curriculum Constructie (meerjarige) Beroepsopleiding tot Huisarts (CCBOH) Rapporten CCBOH 7-13. Utrecht CCBOH, 1986.
5. Harden R, Stevenson M, Wilson Downie W, Wilson GM. Assessment of Clinical Competence using Objective Structured Examination. BMJ 1975;1,447-51.
6. van Leeuwen YD, Pollemans MC, Mol SS, Eekhof JA, Grol R, Drop MJ. The Dutch knowledge test for general practice: issues of validity. Eur J Gen Pract 1995;1:113-7.
7. Kooij LR. De beroepsopleiding tot huisarts. Ervaringen met de nieuwe toelatingsprocedure. Med Cont 1993;48:20-2.
8. Pollemans MC, Tan LHC. Toetsing van kwaliteit, landelijke evaluatie van de interim beroepsopleiding tot huisarts. Rapport SV-1OH-15. Utrecht: SV-IOH; 1990.
9. Wigersma L, Almekinders F, Kooij LR. Raamplan curriculum driejarige huisartsopleiding. Amsterdam/ Utrecht: 1994.
10. Werkgroep Kwaliteitsbewaking. Kwaliteitsbewaking van de individuele huisarts-in-opleiding. Utrecht: SVUH; 1995.
11. Frank JR. The CanMEDS 2005 Physician Competency Framework. Better standards. Better physicians. Better care. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2005.
12. Berkestijn LGM, Wiegersma L, Giesen P, Stalman W. Eindtermen huisartsenopleiding gereed. Med Cont. 2000;55:1234-6.
13. Berkestijn van LGM, Duin BJ van, Hoekstra M, Maiburg HJS, Oosterling EMP, Sagasser MH, Schuling J, Wieringa de Waard M. Competentieprofiel van de Huisarts. Utrecht: NHG; 2005.
14. Protocol Toetsen en Beoordeling in de Huisartsopleiding; PVH 2005.
15. Van der Vleuten CPM, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005;39:309-17.
16. Grol RPTM. Kwaliteitsbewaking in de Huisartsgeneeskunde: effecten van onderlinge toetsing [dissertatie]. Nijmegen: Katholieke Universiteit Nijmegen, 1987.
17. Tan LHC. Tekorten in de opleiding van huisartsen; ziektebeelden en medisch-technische vaardigheden [dissertatie]. Amsterdam, 1989.
18. Pieters HM. De Utrechtse Consult Evaluatie Methode. Vaardigheden in consultvoering getoetst [dissertatie]. Utrecht: Universiteit Utrecht, 1991.
19. Pollemans M. Kennistoetsing bij huisartsen [dissertatie]. Maastricht: Universitaire Pers Maastricht, 1994.
20. Van Leeuwen YD. Growth in knowledge of trainees in general practice [dissertation]. Maastricht: Universitaire Pers Maastricht, 1995.
21. Jansen K. Toetsing van technische vaardigheden van huisartsen. Studies naar toepassingsmogelijkheden van vaardigheidstoetsing in deskundigheidsbevordering [dissertatie]. Unigraphic Maastricht, 1998.
22. Ram P. Comprehensive assessment of general practitioners; a study on validity, reliability and feasibility [dissertation]. Unigraphic Maastricht, 1998.
23. Kramer AJ. Acquisition of clinical competence during postgraduate training in general practice [dissertation]. Universitaire Pers Maastricht, 2003.
24. Pelgrim EAM. Clarifying observation and assessment feedback in workplace-based learning [dissertation]. UMC St Radboud Nijmegen, 2013.
25. Kuyvenhoven MM, Vermeulen MI, van Campen SM, Schmidt JE. Veel aiossen haken af. Med Cont. 2010;65:2206-7.
26. Roberts NK, Williams RG. The hidden costs of failing to fail residents. J Grad Med Educ 2011;3:127-9.
27. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. Acad Med 2005; 80 (10 suppl):S84-S87.
28. Project Vernieuwing Huisartsopleiding. Almekinders WR, Vogeler MC. Deelproject Uitbreiding Capaciteit & Instroom: onderdeel instroom, 2004.

29. Siu E, Reiter HI. Overview: What' s worked and what hasn't as a guide towards predictive admissions tool development. Adv in Health Sci Educ. 2009;14:758-75.
30. Prideaux D, Roberts C, Eva K, Centeno A, McCrorie P, McManus C, et al. Assessment for selection for the health care professions and specialty training: Consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach 2011;33:215-23.
31. Schmidt FL, Hunter JE. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychol Bull 1998;124:262-74.
32. Robertson IT, Smit M. Personnel selection. J Occup Organ Psychol 2001;74:441-72.
33. Wood TJ. Exploring the role of first impressions in rater-based assessments. Adv in Health Sci Educ 2013; doi10.1007/s10459-013-9453-9.
34. Palmer JK, Loveland JM. The influence of group discussion on performance judgments: rating accuracy, contrast effects and halo. J Psychol 2008;142:117-30.
35. Gingerich A, Regehr G, Eva KW, Rater-based assessments as social judgements: rethinking the etiology of raters errors. Acad Med 2011;86:S1-7.
36. HVRC. Model Reglement Selectiecommissies. May 2007.
37. Katzenbauer M. "Die cultuur paste niet bij mij". Med Cont. 2009;64:580-3.

Chapter 2

Selection for Dutch postgraduate GP training; time for improvement

Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, Tromp F, van der Graaf Y, Pieters HM

Eur J Gen Pract 2012;18:201-5.

Abstract

Background

In the Netherlands we select candidates for the postgraduate GP training by assessing personal qualities in interviews. Because of differences in the ratio of number of candidates and number of vacancies between the eight departments of GP training we questioned whether the risk of being rejected diverged amongst them.

Objective

The research question of this study was to which degree department of choice, candidates' characteristics and qualities assessed during interviews explain admission into GP training.

Methods

A nationwide observational study was conducted of all candidates who applied for postgraduate GP training in 2009/2010. Application ratio per department, candidates' characteristics (gender, age, region of medical school and times of application) and qualities (motivation, orientation on the job, personal attributes and learning needs) were collected. Outcome measures were admission to interview and admission to GP training.

Results

The study population addressed 542 candidates. Sixty three candidates were rejected on application letter (11.6%). So 479 candidates were admitted to the interview, of which 340 were admitted to the GP training (71%). Gender and region of medical school outside north western Europe were associated with admission to the interview. Department of choice had a strong association with admission in both stages (RR: 0.30 – 0.74; 0.20 – 0.79 respectively), while candidates' qualities explained admission (RR: 1.09 – 1.25) as well.

Conclusion

The influence of department of choice yields doubts about fairness of the procedure. So advantages and disadvantages of a national procedure are discussed as well as those of a competency based procedure.

Background

Historically, medical doctors in European countries were allowed to work as a General Practitioner (GP) or family doctor after their formal registration as MD. With the emergence of postgraduate GP training in the latter decennia of the 20th century selection procedures were developed in order to improve the quality of physicians in primary care. Selection procedures vary substantially in Europe, however (Table 1) some countries (e.g. Austria and Finland) check minimum criteria as formal registration as MD. Most countries, eg. Iceland, Sweden and the Netherlands, carry out a selection procedure driven by disciplined based training; they aim to select best doctors by assessments of medical knowledge and personal qualities.[1] These personal qualities, considered suitable for completing the postgraduate GP training, are assessed by interviews.

In general the reliability of interview methods to assess candidates' personal qualities is moderate to sufficient with two or more well-trained interviewers.[2] However, the predictive validity of personal interviews with a view on academic and clinical performance is equivocal.[3-5] Rather recently, the United Kingdom has introduced a competency-based selection after an extended job-analysis. Six competencies were targeted for selection: empathy and sensitivity, communication skills, clinical expertise, problem solving, professional integrity and coping with pressure.[6] In Denmark a competency based selection procedure containing Multiple Mini Interviews (MMIs) has been introduced. MMIs apply principles of OSCE to the interview context, which have shown good predictive correlations with future performance.[7,8]

In the Netherlands, the three-year postgraduate GP trainings follow national regulations on selection with personal quality assessments by interviews being the core of the selection procedure.[9] However, there are differences in the number of candidates in comparison with the number of vacancies amongst the eight training departments. Therefore, we questioned whether such a decentralized selection procedure is fair to candidates. The aim of this study was to investigate to which degree the department of choice, and candidates' individual characteristics and qualities explain admission to GP training. The study addressed routinely registered data; so the factual procedure was investigated.

Methods

Design

A nationwide observational study, in cooperation with all eight departments, of all candidates who applied for the GP training was conducted in October 2009 in 7 departments, and for the remaining department in April 2010. Candidates were registered centrally to avoid double applications; the selection procedure was carried out decentralized in the department of candidates' choice. Four departments followed an informed consent procedure.

Selection procedure

First, the local selection committees decide which candidates are invited to the interview stage using criteria as mastery of the Dutch language and quality of motivation, expressed in their respective letters of application. Second, three members of the selection committee, staff member, GP trainer, GP trainee, independently assess candidates' qualities as motivation, orientation on the job, personal attributes and learning needs by means of semi-structured interviews (30-45 minutes). Relevance of curriculum vitae and candidates' own perception of being a future GP, are somewhat differently assessed by the departments. The final conclusion on admission into GP training is based on a consensus procedure with the selection committees considering all aspects.

Data collection

Number of candidates, number of vacancies per department and individual characteristics were derived from the national Dutch postgraduate GP training (Huisarts Opleiding Nederland). All data were clerically depersonalized before data processing.

Individual characteristics contain age when selected (in years), gender (male versus female), region of medical school (NW Europe versus elsewhere) and the number of times of application (first time versus second time or more). Candidates' qualities (motivation, orientation on the job, personal attributes and learning needs) were assessed by the three members of the selection committee on ordinal scales with two extremes (poor/ insufficient and very good/excellent), varying from 3 points to 10 points at the respective departments. The three independent scores per quality were averaged.

Outcome measures were admission to the interview (stage 1) and admission to the GP training (stage 2).

Analysis

The associations between the determinants and admission to the interview and admission to the GP training were estimated by means of relative risks (RRs) with Log binomial models, followed by subgroup analysis.[10] As there was a difference between number of applicants and number of vacancies per department, we calculated the application ratio and included this as an offset in the model to correct for these differences. The department with the lowest application ratio was used as a reference group. By using Z scores results of different rating scales could be compared. All analyses were done in SAS (version 9.1).

Results

In total, 597 candidates applied for the selection procedure and 375 (62.8%) were admitted to the postgraduate GP training (Table 2). Nine per cent of the candidates (55/597) were not evaluable, mostly because of lack of informed consent. Therefore, the study population consisted of 542 candidates. One third of the candidates was male (*n* = 181, 33.4%). Mean age was 29.9 years (SD 5.2, minimum 22, maximum 52 years old), 506 (93.4%) followed their medical school in north western Europe and 392 (72.6%) applied for the first time. 63 were rejected on application letter (11.6%). Therefore, 479 candidates were admitted to the interview, of which 340 were eventually admitted to the GP training (71%).

In the first stage, 88% of the candidates were selected for the interview stage. Taking the department with the lowest application ratio as a reference there was an independent association between the department of choice and admission to the interview, which means that the more candidates per vacancies, the less probability to be admitted to the interview stage (Table 3). Male candidates and those who followed their medical education outside north western Europe had a smaller probability of admission to the interview too.

In the second stage, the probability of eventual admission was strongly related to the application ratio too. Candidates for departments 4, 5, 6, 7 and 8 had a far more low risk (RR; 0.20 – 0.45; Table 3) to be admitted than those for the reference. In addition higher ratings on motivation, orientation on the job and personal attributes explained independently eventual admission; these findings were the case for all eight departments.

Discussion

Summary of main finding

Department of choice was a strong predictor for admission in both stages. Candidates' qualities assessed during the interview explained eventual admission too. The impact of the differences in application ratio between the respective departments of choice yields doubt about fairness and public defensibility of the decentralized Dutch selection procedure.

Strengths and limitations of the study

As a result of strong collaboration between departments, it was possible to collect data of selection procedures in 2009/2010. Informed consent procedure at four departments has led to some missing data; selection bias might be possible. However we assume low bias on observed associations, because results of departments with and without an informed consent procedure did not differ. Data on assessors were not available, so we could not check possible assessor bias.

Interpretation

Gender and region of medical school outside north western Europe was associated with admission to the interview. The association with gender is probably caused by the fact that women express their motivation better than male. The impact of the region of medical school might be caused by insufficient mastery of Dutch language. Personal attributes, motivation and orientation on the job were determinants of eventual admission to the GP training, which is in line with the aim of the procedure. The department of choice is the strongest predictor for admission. This might be caused by the fact that the respective departments practice relative criteria partly based on the ratio between candidates and vacancies. So candidates who have been rejected in one department might have been admitted elsewhere and vice versa. This latter result yields doubts about fairness of the current procedure.

Local versus national procedure

A decentralized procedure as the Dutch seems to cause inequality and unfairness to the candidates, which can be hardly prevented by training of local selection committees. A national procedure with assessors blinded for selection on behalf of their own department could resolve this problem. While fairness to candidates and public defensibility are important advantages of such a national selection procedure, there are disadvantages too. A national selection procedure will lead to less involvement of local staff and trainers in selection of their own trainees. In addition, such a national selection procedure will yield fewer possibilities to develop local identity of departments by preference of type of candidates.

Discipline versus competency based

The reconstruction of a historically more discipline based Dutch GP training into a competency based curriculum in 2005 and the doubtful predictive validity of discipline based selection procedures argue for an improvement of the procedure. Provisional results from the UK are promising in showing that trainees recruited by means of competency based methods performed better on key competencies after three months in practice than those recruited through traditional selection procedures.[11] MMI's are chosen in Denmark to select trainees on clear competency based criteria. These are nowadays considered rather appropriate for such an aim.[7,8]

The doubtful predictive validity of the current discipline based procedure and the promising results mentioned above advocate a competency-based-selection procedure, which has several advantages. First, it gives the opportunity to assess candidates on relevant competencies. Second, incompetent candidates can be selected out on rather clear arguments. Last, competency based selection can provide an individual training/ educational plan for admitted trainees. The individual educational plan can be assessed during the GP training, which is a good preparation for future monitoring of functioning as GP. How far competency based selection procedures is more time consuming and so more expensive than discipline based procedures depends on a lot of factors such as costs of baseline situation, number of instruments and number of assessors.[15] Using machine-marked tests (like knowledge and situational judgement tests) for pre-selection might reduce costs.

Conclusion

Although poor performance and attrition rate of GP trainees in the Netherlands is relatively low, we have public responsibility to perform a fair and defensible selection procedure.[12] With a view on the results of this study, we would plead for a national procedure. In addition, with the reconstruction into a competency-based curriculum in 2005 and the promising results of competency based selection, the Dutch GP training has to commit itself to a competency based selection procedure based on critical GP competencies.[1,13] The selection procedure has to be part of a more complex strategy to strengthen primary care and prevent wasting capacity of competent doctors.

References

1. Sammut MR, Lindh M, Rindlisbacher B & on behalf of EURACT- the European Academy of Teachers in General Practice. Funding of vocational training programmes for general practice/ family medicine in Europe. Eur J Gen Pract. 2008;14:83-8.
2. Kreiter CD, Yin P, Solow C, Brennan RL. Investigating the reliability of the medical school admissions interviews. Adv in Health Sci Educ 2004;9:147-59.
3. Albanese M, Snow M, Skochelak S, Huggett K, Farrell P. Assessing personal qualities in medical school admissions. Acad Med 2003;78:313-21.
4. Salvatori P. Reliability and validity of admission tools used to select students for the health professions. Adv in Health Sci Educ 2001;6:159-75.
5. Siu E, Reiter HI. Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. Adv in Health Sci Educ 2009;14:758-75.
6. Patterson F, Ferguson E, Lane PW, Farrell K, Martlew J, Wells AA. Competency model for general practice: implications for selection, training and development.Br J Gen Pract 2000;50:188-93.
7. Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. Predictive validity of the multiple mini-interview for selecting medical trainees. Med Educ 2009;43:767-75.
8. Prideaux D, Roberts C, Eva K, Centeno A, McCrorie P, McManus C, Patterson F, Powis D, Tekian A, Wilkinson D. Assessment for selection for the health care professions and specialty training: Consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach 2011;33:215-23.
9. Regulations selection committees HVRC (in Dutch: Model Reglement Selectiecommissies) HVRC May 2007.
10. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. Am J Epidemiol 2003;157:940-3.
11. Patterson F, Ferguson E, Norfolk T, Lane P. A new selection system to recruit general practice registrars: preliminary findings from a validation study. BMJ 2005;330:711-4.
12. Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, Graaf van der Y, Pieters HM. Poor performance and attrition of the residents in first year of GP training. Ned Tijdschr Geneeskd 2011;155: A2780.
13. Baarveld F, Bottema BJAM, Bueving HJ, Langendoen-Roel M, Leeuwen van YD, Pieters HM, Schoonheim PL, Wieringa- de Waard M. Raamcurriculum 2005 (in Dutch) SVUH0502.
14. Patterson F, Denneey ML, Wakeford R, Good D. Fair and equal assessment in postgraduate training? A future research agenda. Br J Gen Pract. 2011;61:712-3.
15. Rosenfeld J, Reiter H, Trinh K, Eva K. A cost efficiency comparison between the multiple mini – interview and traditional admission interviews. Adv in Health Sci Educ 2008;13:43-58.

Table 1 European overview of entrance procedure for post graduate GP training. Based on correspondence with key persons of EURACT.[a]

Registration and Selection procedure			
Registration **Checking minimum criteria: MD**	**Discipline based assessment procedure**		**Competency based assessment procedure**
No additional criteria	*In case of sufficient vacations, all candidates are admitted*	*Threshold scores on knowledge and or personal qualities*	*Threshold scores on competencies (and knowledge tests)*
Austria, Bosnia & Herzegovina, Croatia, Finland, Flanders, Greece, Montenegro (>2 yr MD) Norway, Slovakia	Czech, Latvia, Poland, Portugal, Slovenia, Switzerland	Cyprus, Estonia, Hungary, Iceland, Ireland, Italy Netherlands, Romania, Spain, Sweden	Denmark, United Kingdom

[a] *Unknown or no GP certification: Albania, Bulgaria, France, Germany, Lithuania, Serbia, Wallonia*

Table 2 Number of candidates, vacancies, admission and evaluable candidates, baseline characteristics of evaluable candidates per department.

Department	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**
Number of candidates	27	34	38	62	66	134	155	81	597
Number of vacancies	36	36	36	36	36	72	80	36	368
Ratio candidate/ vacancy	0.75	0.94	1.03	1.72	1.83	1.86	1.91	2.25	1.62
Number admitted	23	27	28	44	46	80	87	40	375
Percentage admitted	85.2	79.4	73.7	71.0	69.7	59.7	56.1	49.4	62.8
Number evaluable	26[a]	24[b]	37[c]	29[d]	64[e]	131[f]	150[g]	81	542
Male, *n* (%)	11 (42.3)	9 (37.5)	17 (45.9)	12 (41.4)	15 (23.4)	42 (32.1)	53 (35.3)	22 (27.2)	181 (33.4)
Age in years mean (SD)	32.0 (8.5)	32.1 (7.7)	29.8 (5.5)	28.2 (4.8)	28.8 (4.8)	30.3 (5.0)	29.9 (4.6)	29.0 (4.2)	29.9 (5.2)
Medical school NW Europe, *n* (%)	25 (96.2)	19 (79.2)	33 (89.2)	27 (93.1)	59 (92.2)	125 (95.4)	140 (93.3)	78 (96.3)	506 (93.4)
First time of application, *n* (%)	19 (73.1)	12 (50.0)	30 (81.1)	22 (75.9)	54 (84.4)	84 (64.1)	108 (72.0)	63 (77.8)	392 (72.6)

[a] Withdrawal during procedure: 1, [b] No informed consent: 10, [c] No informed consent: 1, [d] No informed consent: 33, [e] Withdrawal during procedure: 2, [f] Withdrawal during procedure: 3, [g] Missing 3, no informed consent: 2

Table 3 Univariate and Multivariate Relative Risks (RR (95% CI)) of being admitted to the interview stage (Stage 1) and of being admitted to the GP training (Stage 2)

	Admission to interview Stage 1 *n* = 542		**Admission to GP training Stage 2 *n* = 479**	
	Univar. RR (95% CI)	**Multiv. RR (95% CI)**	**Univ. RR (95% CI)**	**Multiv. RR (95% CI)**
Department 1 (= reference)	1.00	1.00	1.00	1.00
Department 2	0.69 (0.60 – 0.81)	0.74 (0.63 – 0.87)	0.77 (0.62 – 0.96)	0.79 (0.65 – 0.97)
Department 3	0.63 (0.57 – 0.71)	0.63 (0.57 – 0.70)	0.68 (0.56 – 0.83)	0.68 (0.56 – 0.81)
Department 4	0.38 (0.32 – 0.43)	0.36 (0.32 – 0.42)	0.43 (0.35 – 0.53)	0.45 (0.36 – 0.56)
Department 5	0.40 (0.38 – 0.41)	0.38 (0.35 – 0.41)	0.34 (0.28 – 0.42)	0.34 (0.29 – 0.41)
Department 6	0.33 (0.30 – 0.36)	0.32 (0.29 – 0.36)	0.34 (0.29 – 0.41)	0.34 (0.29 – 0.40)
Department 7	0.33 (0.31 – 0.36)	0.33 (0.30 – 0.36)	0.30 (0.25 – 0.36)	0.29 (0.25 – 0.35)
Department 8	0.32 (0.30 – 0.33)	0.30 (0.28 – 0.33)	0.20 (0.15 – 0.25)	0.20 (0.16 – 0.24)
Gender (male = ref)	1.11 (1.01 – 1.21)	1.08 (1.01 – 1.16)	1.00 (0.87 – 1.16)	0.95 (0.86 – 1.06)
Age (in years)	0.98 (0.97 – 0.99)	0.99 (0.98 – 1.00)	0.98 (0.96 – 1.00)	1.00 (0.99 – 1.01)
Region med school (NW Eur = ref)	0.63 (0.45 – 0.88)	0.66 (0.48 – 0.90)	0.48 (0.24 – 0.95)	0.68 (0.37 – 1.27)
Times of application (1st time = ref)	0.94 (0.89 – 1.00)	0.96 (0.92 – 1.02)	0.94 (0.84 – 1.04)	1.00 (0.93 – 1.08)
Motivation			1.51 (1.39 – 1.63)	1.20 (1.10 – 1.31)
Orientation on the job			1.41 (1.31 – 1.52)	1.09 (1.02 – 1.17)
Personal attributes			1.54 (1.42 – 1.67)	1.25 (1.14 – 1.36)
Learning needs			1.49 (1.38 – 1.61)	1.09 (0.99 – 1.19)

Chapter 3

Dutch postgraduate GP selection procedure; reliability of interview assessments

Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, van der Graaf Y, Damoiseaux RAMJ

BMC Fam Pract 2013,14:43 doi: 10.1186/1471-2296-14-43.

Abstract

Background

Semi-structured interviews are the core of the Dutch selection procedure for postgraduate general practice (GP) training. A staff member, trainer and trainee independently assess personal qualities. Aiming to improve the selection procedure we were interested in the reliability aspects of these interviews. We investigated the inter-rater reliability of the interview for groups of two or three assessors and the degree to which candidates' characteristics and qualities assessed during interviews explained admission into GP training, controlled for differences between those who apply for the first versus the second or third application.

Methods

An observational study was conducted of all candidates who entered the Utrecht selection procedure between April 2008 and 2010. Candidates' characteristics and qualities were collected. Inter-rater reliability of different compositions of the interview group per quality was estimated. Factors associated with admission into GP training were assessed.

Results

The study population included 394 candidates. Twenty-six candidates were rejected based on their application letter (4.4%). Three candidates who applied more than 3 times were excluded. Ultimately, 206 of the 365 candidates were admitted to the GP training (56.4%). The inter-rater reliability was satisfactory (ICC: 0.78 – 0.84). Reduction from three to two assessors slightly reduces the ICC. The candidates' qualities independently explained admission to GP training, whereas individual characteristics did not. These results did not differ for candidates who applied for the first time versus candidates applying for the second or third time.

Conclusion

Selection interviews with two assessors yielded a satisfactory level of reliability. Individual characteristics were not associated with admission, whereas scores related to candidate qualities did show such an association. The results of those applying for the second or third time were similar.

Background

The core of the present Dutch selection procedure for postgraduate general practice (GP) training includes semi-structured interviews. These personal interviews are conducted by a staff member, a trainer and a trainee to assess candidates' motivation, orientation on the job, learning needs and personal attributes. Comparable selection methods are used in many European countries, originating from the discipline based training model developed in the last quarter of the 20th century.[1] In general, the reliability of interview assessments in medical school admission is considered moderate to good. Reliability increases by structuring interviews, training assessors and increasing the number of assessors or interviews.[2-5]

However, this assessment method can be criticised from different points of view. First, it weakly predicts future clinical and academic performance.[3,4,6,7] In addition, we have recently found that the current interview procedure yields doubts about fairness for candidates, and the respective departments of choice have a strong influence on admission.[1] Given these considerations, the national Dutch GP training (Huisarts Opleiding Nederland) aims to update the selection to a competency- based procedure with an extension of instruments.[8] As we decided to maintain a highly structured interview in the new procedure, we investigated the reliability of interview assessments in the current procedure with three groups of assessors (staff members, trainers and trainees). From an economic perspective, we explored the degree to which reliability diminishes in case of reduction from three to two interview assessors. Another aim of this study was to determine whether our earlier findings, that individual characteristics such as age and gender, do not predict admission into GP training, could be replicated. In addition, we explored whether the results differed for candidates who applied for the first time versus the second or third time.[1] The data for this study are the routinely registered data of the selection procedure on the department of Utrecht from 2008 – 2010.

Methods

Design

An observational study of all candidates who entered the Utrecht selection procedure between April 2008 and April 2010 was conducted.

Selection procedure

After national registration, the selection for Dutch GP training is conducted locally at the department of each candidate's choice. The local selection committee decides which candidates are invited to the interview using criteria such as mastery of the Dutch language and the quality of motivation expressed in their letters of application. Each member of the selection committee, which consists of a staff member, a GP trainer and a

GP trainee, independently assesses the qualities of the candidates, including their motivation, orientation on the job, learning needs and personal attributes, after a personal interview with a duration of 30 to 45 minutes.[1] All assessors receive written and oral training at the beginning of the selection procedure to learn how to question and score these qualities.

Box 1 Sample questions from the semi-structured interview

Motivation to become a GP:
Why did you choose to become a GP among all of the specialisations?
Did you consider other specialisations?
What type of GP do you want to become in the future?
What is the relevance of your CV?

Learning needs/learning styles of candidates:
What are your strengths and weaknesses in learning?
What methods are helpful for you in developing your knowledge, skills and attitude?
What is your experience with group sessions, video assessments, OSCEs and other activities in relation to your own learning?

Orientation/insight on the job as a GP:
What do you know about the range of tasks/job responsibilities of a GP?
What do you know about collaboration with other disciplines?
What medical journals did you read to prepare for postgraduate training?
What is your future vision as a GP?

Personal attributes in relation to clinical performance:
How do you make decisions?
How do you take responsibility?
How do you cope with pressure and uncertainty?
How do you provide and handle feedback?

Data collection

All data were derived from the Utrecht postgraduate GP training. Ethical approval for routinely gathered data was not mandatory at the time this study was conducted. Therefore, we executed the study according to the 'code of conduct' for the use of personal data in scientific research. Before data processing, all data were clerically anonymised.

Individual characteristics were age at the moment of selection (in years); gender (male versus female); region of medical school (NW Europe versus elsewhere); past clinical performance after graduation (less than one year; more than one year) and the number of times of application (first time versus second or third time). Candidates' qualities (motivation, orientation on the job, learning needs and personal attributes) were independently rated on a three point scale by the three members of the selection committee

- below standard (1), standard (2), above standard (3).
The outcome measure was: admission into postgraduate GP training.

Analysis

We first explored differences between the characteristics and qualities of candidates who applied for the first time versus those who applied for the second or third time. Subsequently, we described reliability aspects, with mean quality scores (SD) according to the three groups of assessors. Inter-rater reliability was estimated for each quality with intraclass correlation coefficients (ICC), calculated for all assessors and any combination of two assessors.[9] Associations between the characteristics and qualities and admission into postgraduate training were estimated with log binomial models. Therefore they are reported as relative risks.[10] In case of missing data (1.3% of data values), mean values or modal category scores were imputed.[11]

Some candidates (*n* = 50) were included more than once in our study population due to consecutive selection procedures. We thus controlled whether the association between determinants and the outcome differed for those who applied for the first time (model 1) versus those who applied for the second or third time (model 2). We computed the linear predictor for candidates who applied for the second and the third time based on the analysis of candidates who applied for the first time.[12-14] This linear predictor was subsequently analysed as a single determinant in model 2. If the results from model 1 are valid for second and third time candidates, the regression coefficient of this linear predictor in model 2 will be close to 1. The analysis was done in SPSS version 17 and SAS version 9.2.

Results

Candidates' characteristics

Three hundred ninety four candidates applied for the postgraduate GP training between April 2008 and April 2010 in Utrecht. Twenty-six were rejected based on their letter of application. Candidates who applied more than 3 times were excluded (*n* = 3). A total of 365 candidates were included in the study population: 264 applied for the first time, 87 for the second time and 14 for the third time. One fourth of the candidates were male, the mean age was 29.7 years (SD 4.9) and 94.5% followed medical school in north western Europe (Table 1). The group who applied for the second or third time was older and had more clinical experience. The mean score of the candidates' qualities varied from 2.0 (orientation on the job) to 2.3 (motivation). Candidates who applied for the second or third time had approximately the same scores on personal qualities as those who applied for the first time, with one exception: they had lower scores on personal attributes (Table 1).

Reliability

There were almost no differences in mean scores between the three groups of assessors, or in the standard deviation (Table 2). The reliability of the scores among assessors was good, with the lowest score for learning needs (ICC: 0.78 – 0.84). If the assessments of trainees were deleted, the ICC diminished least (ICC: 0.73 – 0.79). The reduction of the ICC was highest (ICC: 0.68 – 0.75) in case of deleting the assessments of the group of staff members. There were no differences regarding reliability between candidates who applied for the first time versus the second or third time (not shown in a table). There was a moderate to strong association amongst the four qualities (Pearson's r: 0.40 – 0.64, not shown in a table), indicating that those who scored rather high on motivation did also on orientation on the job, learning needs and personal attributes and vice versa.

Predictors

Each of the four candidates' qualities was independently associated with being admitted into the GP training (Table 3), with personal attributes and motivation being the strongest predictors. Individual characteristics, such as age and gender, did not show an association with being admitted. We applied the results of the regression analysis of the first application to the candidates who applied for the second or third time, which resulted in a regression coefficient of 0.93 (95% CI 0.72 – 1.15). Therefore, the results in both groups were similar.

Discussion

Summary of main findings

The mean scores and variations in personal qualities awarded by staff members, GP trainers and trainees were nearly the same. The reliability of interview assessments among the three assessors was satisfactory. Exclusion of the assessments of one group (staff member, trainer or trainee) just slightly reduced reliability. Our results show an independent relation between personal qualities, selection criteria, and admission into the postgraduate training; age and gender did not influence the decision.

Discussion of results

Reviews have shown varying reliability in medical school admission interviews, as previous studies were not primarily designed to investigate reliability, because the format and structure of the interview widely vary and because of assessor bias.[2,4,15] The current study demonstrates a satisfactory level of reliability of the candidates' quality assessments, which corresponds with more recent studies.[16,17] This may be an effect of structuring the interview and training the assessors, which are factors known to enhance reliability.[3-5,15] The reliability of the interview assessments in this study can be considered satisfactory as well with a view on the duration of the interviews, because reliability of an assessment procedure partly depends on the duration of the procedure.[18]

At this time the selection committee consists of three groups of assessors, who conduct assessments from their specific perspectives. Our results show that two assessors would have been sufficient in terms of reliability. This finding is in accordance with other studies that find satisfactory reliability between 2 assessors.[19,20] In general the staff member has the most experience in assessing candidates. This is reflected by the somewhat higher ICC's of all pairs of assessments in which the staff member' assessments were included. Extension of the number of instruments, with Multiple Mini Interview (MMI's) regarding to collaboration, professionalism and doctor patient encounters, and further structuring the interviews, may improve reliability.[21,22]

In accordance with our earlier findings candidates' qualities, such as motivation, orientation on the job, and personal attributes, were independently associated with being admitted.[1] Individual characteristics, such as age and gender did not correlate with the decision of being admitted. These findings are in line with the formal procedure and study by Lumb et al.[1,16], whereas Shaw et al. found that the gender and race of candidates influenced the interview scoring.[23]

Strengths and limitations of the study

By using data on five consecutive selection procedures, it was possible to analyse the assessments of more than 300 candidates. The extent of the group made it possible to determine whether the results differed for candidates who applied for the first time versus candidates applying for the second or third time.[1]

This study also has certain limitations. First, the favourable reliabilities may be partly caused by the limited scale width (a three-point scale). However, the literature indicates that the reliability of ratings at the high or low ends of a rating scale is higher than that for the middle levels. Thus, a three-point scale may be as useful as the commonly used five-point scale.[8,24] Controlling the results by calculating the (nonparametric) Kendall's coefficient of concordance W for the candidate quality assessments yielded similar results (coefficient of concordance W for three of the four qualities between 0.75 and 0.77; learning needs: 0.69; all $p < 0.05$).[25]

Secondly, the correlation between qualities may suggest a halo effect, but this cannot be studied further with these data. The candidates were assessed by various assessors. Therefore, the design did not allow a generalisability analysis, nor did the design provide the opportunity to investigate assessment bias by calculating sources of variance.

Conclusion

Interview assessments by two representatives of relevant professional groups – a staff member and a trainer – show satisfactory reliability compared with interviews by three representatives. Given this finding and the promising results from the literature of multiple independent assessments in the selection procedures, we plead for a reduction of the number of assessors in the interviews and an extension of the instruments, eg. with MMI's, for a more reliable and valid competence based procedure.[8]

References

1. Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, Tromp F, van der Graaf Y, Pieters HM. Selection for Dutch postgraduate GP training; time for improvement. Eur J Gen Pract 2012; 18:201–205.
2. Kreiter CD, Yin P, Solow C, Brennan RL. Investigating the reliability of the medical school admissions interviews. Adv in Health Sci Educ 2004;9:147–159.
3. Albanese M, Snow M, Skochelak S, Huggett K, Farrell P. Assessing personal qualities in medical school admissions. Acad Med 2003;78:313–321.
4. Salvatori P. Reliability and validity of admissions tools used to select students for the health professions. Adv in Health Sci Educ 2001;6:159–175.
5. Morris JG. The value and role of the interview in the student admission process: a review. Med Teach 1999; 21:473–481.
6. Siu E, Reiter HI. Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. Adv in Health Sci Educ 2009,14:758–775.
7. Goho J, Blackman A. The effectiveness of academic admission interviews: an exploratory meta analysis. Med Teach 2006;28:335–340.
8. Prideaux D, Roberts C, Eva K, Centeno A, McCrorie P, McManus C, Patterson F, Powis D, Tekian A, Wilkinson D. Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 conference. Med Teach 2011;33:215–223.
9. McGraw KO, Wong SP. Forming inferences about some intraclass correlations coefficients. Psychological Methods 1996;1:30–46.
10. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. Am J Epidemiol 2003;157:940–943.
11. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J of Clin Epidemiol 2006;59:1087–1091.
12. Harrell FE. Regression modelling strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, USA: New York Springer; 2001.
13. Cox DR: Two further applications of a model for binary regression. Biometrika 1958;45:562–565.
14. Steyerberg EW, Eijkemans MJC, Harell FE, Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Statist Med 2000;19:1059–1079.
15. Edwards JC, Johnson EK, Molidor JB. The interview in the admission process. Acad Med 1990;65:167–177.
16. Lumb AB, Homer M, Miller A. Equity in interviews: do personal characteristics impact on admission interview scores? Med Educ 2010;44:1077–1083.
17. Rao R. The structured clinically relevant interview for psychiatrist in training (SCRIPT): a new standardized assessment tool for recruitment in the UK. Acad Psychiatry 2007; 31:443–446.
18. van der Vleuten CPM, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005;39:309–317.
19. Hamel P, Boisjoly H, Corriveau C, Fallaha N, Lahoud S, Luneau K, Olivier S, Rouleau J, Toffoli D. Using the CanMEDS roles when interviewing for an ophthalmology residency program. Can J Ophthalmol 2007;42:299–304.
20. Patrick LE, Altmaier EM, Kuperman S, Ugolini K. A structure interview for medical school admissions, phase 1: initial procedure and results. Acad Med 2001;76:66–71.
21. Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. Predictive validity of the multiple mini-interview for selecting medical trainees. Med Educ 2009;43:767–775.
22. Patterson F, Ferguson E, Norfolk T, Lane P. A new selection system to recruit general practice registrars: preliminary findings from a validation study. BMJ 2005;330:711–714.
23. Shaw DL, Martz DM, Lancaster CJ, Sade RM. Influence of medical school applicants' demographic and cognitive characteristics on interviewers' ratings of noncognitive traits. Acad Med 1995;70:532–536.
24. Stansfield R, Kreiter C. Conditional reliability of ratings: extreme ratings are the most informative. Med Educ 2007;41:32–38.
25. Siegel S, Castellan NJ. Nonparametric statistics. 2nd edition. New York, USA: New York McGraw-Hill Book Company; 1988;262–272.

Table 1 Baseline characteristics of candidates

Individual characteristics	1^{st} time application *n* = 264	$2^{nd}/3^{rd}$ time application *n* = 101	Total *n* = 365
Gender male, *n* (%)	68 (25.8)	30 (29.7)	98 (26.8)
Age, mean in years (SD)	29.2 (4.7)	31.0 (5.2)	29.7 (4.9)
Medical school NW Europe, *n* (%)	251 (95.1)	94 (93.1)	345 (94.5)
Past clinical performance < 1 year, *n* (%)	136 (51.5)	34 (33.7)	170 (46.6)
Candidates' qualities			
Motivation, total mean score (SD)	2.3 (0.6)	2.4 (0.5)	2.3 (0.6)
Orientation on the job, total mean score (SD)	2.0 (0.5)	2.1 (0.5)	2.0 (0.5)
Learning needs, total mean score (SD)	2.3 (0.5)	2.2 (0.6)	2.2 (0.5)
Personal attributes, total mean score (SD)*	2.3 (0.6)	2.1 (0.5)	2.2 (0.6)
Admitted, *n* (%)	148 (56.1)	58 (57.4)	206 (56.4)

SD standard deviation
**difference 0.2 (CI 95%: 0.1 – 0.3)*

Table 2 Mean scores (SD) of the assessed qualities according to interviewer; Interrater reliability (Intraclass correlation, ICC)

Quality *n* = 365	Staff member mean(SD)	Trainer mean(SD)	Trainee mean(SD)	ICC	ICC without staff member	ICC without trainer	ICC without trainee
Motivation	2.3 (0.6)	2.3 (0.7)	2.3 (0.6)	0.84	0.75	0.81	0.79
Orientation on the job	2.0 (0.6)	2.0 (0.6)	2.0 (0.6)	0.84	0.75	0.78	0.79
Learning needs	2.2 (0.6)	2.3 (0.6)	2.2 (0.6)	0.78	0.68	0.68	0.73
Personal attributes	2.2 (0.7)	2.3 (0.6)	2.2 (0.7)	0.83	0.71	0.78	0.79

SD standard deviation

Table 3 Univariate and Multivariate Relative Risks (95% CI) of being admitted to the GP training

***n* = 365**	**Univariate RR (95% CI)**	**Multivariate RR (95% CI)**
Age (in years)	0.95 (0.92 – 0.97)	0.98 (0.96 – 1.00)
Gender (male=ref)	0.87 (0.70 – 1.09)	1.09 (0.92 – 1.28)
Region (NW Europe=ref)	2.33 (1.09 – 5.01)	1.18 (0.65 – 2.14)
Past performance (< 1 year= ref)	1.08 (0.90 – 1.30)	1.12 (0.97 – 1.29)
Motivation	3.18 (2.67 – 3.78)	1.76 (1.46 – 2.12)
Orientation on the job	2.38 (2.03 – 2.79)	1.34 (1.14 – 1.59)
Learning needs	3.20 (2.72 – 3.77)	1.42 (1.17 – 1.73)
Personal attributes	3.09 (2.64 – 3.62)	1.84 (1.54 – 2.19)

CI confidence interval
ref reference group

Chapter 4

Poor performance and attrition among trainees in the first year of their postgraduate GP training

Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, van der Graaf Y, Pieters HM

Ned Tijdschr Geneeskd 2011;155:A2780

Abstract

Objective

To investigate which determinants were related to poor performance and involuntary attrition in the first year of the postgraduate GP training.

Design

Observational cohort study of trainees who started the GP training in Utrecht between 2005 and 2007.

Methods

Individual characteristics (as age, gender), overall scores of the competency roles 'medical expert', 'communicator' and 'professional' and the scores of the National GP Knowledge Test were collected. The outcome measure was: poor performance and involuntary attrition in the first year of the training.

Correlations between assessment scores after three months on the three competency roles and knowledge were measured. Next the association between individual characteristics, early competency assessments and knowledge scores and the outcome measure were estimated by means of logistic regression analysis (adjusted ORs; 95% CI).

Results

Totally, 215 trainees started the GP training. In the first quarter of the year about 25% of the trainees were rated as insufficient on one or more of the roles. Competencies were mutually correlated, but not with the knowledge test. Eighteen trainees exhibited poor performance and three were forced to stop the training. Higher age (adj OR 1.1; 1.0 – 1.3), insufficient assessment score as 'medical expert' (adj OR 2.1; CI 1.1 – 4.0) and knowledge (adj OR 8.9; CI 3.0 – 26.3) were independently correlated with the outcome.

Conclusion

Higher age, insufficient scores on 'medical expert' role and insufficient knowledge at the beginning of the training are risk factors for poor performance and involuntary attrition.

Introduction

In the Netherlands, about one of ten trainees on internal medicine or surgery prematurely leaves the postgraduate training program.[1] Long working hours, hierarchy and monotonous work are mentioned as reasons. Recently, we have shown that the mean attrition rate of the Dutch GP training has been 7.5%, almost half occurs in the first year of the training.[2] The number of trainees who exhibit poor performance is higher.

GP trainers and staff members who advise on the progress to the second year, have the impression that poor performance in the first year, leads to shortcomings in the competency roles as ‘medical expert’, ‘communicator’ and ‘professional’. The first year of the postgraduate GP training primarily focuses on these three roles, which is in line with the Framework Curriculum of General Practice Training 2005.[4] The other roles (‘collaborator’, ‘manager’, ‘health advocate’ and ‘scholar’) receive more attention in the second and third year.

From a policy perspective, it is worrying that a training program stagnates or has to end; for the trainee in question, his or her trainer and staff members, these processes are laborious and time-consuming. Therefore it is important to identify determinants of poor performance and involuntary attrition.

This retrospective observational study describes poor performance among trainees in the first year of the GP training and its associated determinants. The research questions are:

- How many trainees exhibit poor performance or involuntary attrition and in which competency roles do shortcomings appear?
- To what degree are individual characteristics and early competency/knowledge assessments in the first quarter associated with poor performance or involuntary attrition at the end of year one?

Methods

The Utrecht Trainee Competence Study in General Practice (GP-UTCS) is a retrospective observational cohort study of all trainees who started the GP training Utrecht between March 1st 2005 and September 1st 2007.

Data collection

The following data were collected from the dossiers of the trainees:

Individual characteristics: gender; age (in years) at the time of entry into the program; region of medical school (NW Europe; elsewhere); past performance-clinical experience as a medical doctor (< one year or ≥ one year); the number of times the trainee applied to the training program (once; more than once).

Assessments: overall scores on three competency roles (‘medical expert’, ‘communicator’ and ‘professional’) in the first and last quarter of the year; the score of the first

National GP Knowledge Test (LHK, Dutch abbreviation) in the second month of the program (insufficient/sufficient).[5,6]

The competency scores were recorded in the Compass, a competency assessment list (example in Figure 1). This list includes behavioural indicators regarding the mastery of competencies in seven roles, expressed on a four point scale (1 = poor; 2 = insufficient; 3 = sufficient; 4 = good).[5] The reference point was the end of year one. In about one third of the trainees the precursor of the Compass was used, a national form to evaluate the trainee (called LEV list, Dutch abbreviation) with similar items and scale; the reference point was the final level of the GP training.

The LHK is a knowledge progress test with 160 true/false questions which is administered twice a year by all GP trainees. The pass/fail limit is determined on the scores of all trainees in the same phase of training. Trainees who score below the national mean minus 1 SD obtain 'insufficient'.

Outcome measure was: poor performance ('go, unless') and involuntary attrition from the first year ('no go') as decided by the head of the GP training at the end on the first year. This decision is based on the evaluations of the GP trainer and the staff members and the results of the LHK.[7]

Analysis

Individual characteristics and early assessment scores were described. The correlations between the competency scores and the LHK scores at the beginning of the first year were calculated using the Spearman's rank correlation coefficient (rho). In the dossiers of the trainees who received judgement 'go, unless' or 'no go', we investigated in which competency roles the trainees showed their shortcomings, and, whether these shortcomings were reflected in the competency scores in the last quarter of the first year.

Next, the association between individual characteristics, competency and LHK scores (sufficient versus insufficient) in the first quarter and poor performance/involuntary attrition were estimated by multivariate logistic regression analysis (adjusted ORs, 95% CI). First, the association between individual characteristics and the outcome and then the association between early assessment scores and outcome were estimated. Based on these two analyses, the variables with $p \leq 0.05$ and the variables relevant from theoretical perspectives, were included in the final model.

Because two different competency assessment lists (Compass and LEV, same scale, but different reference point) were completed in this cohort z-scores were used, based on the distribution of the scores on the respective assessment lists. In total 4% of the data was missing, most likely by relatively low awareness of the Compass, which at the time of the study was recently introduced. Multiple imputation was applied because missing values may lead to bias.[8-9]

Informed consent

At the time of the study ethical approval was not mandatory for routinely gathered data. We executed the study according to 'the code of conduct' for the use of personal data in scientific research (http://www.vsnu.nl/code-pers-gegevens.html, in Dutch 2005). Before data processing, all data were clerically anonymised.

Results

Between March 1st 2005 and September 1st 2007, 215 trainees started the GP training in Utrecht (Table 1). Two-third of these trainees was female. The mean age at the beginning of the training was 29.5 years [min 24, max 46]. Seven per cent of them had completed their medical school outside north western Europe. More than half of the trainees had more than one year clinical experience as MD at the time of application and one-fourth of the trainees previously applied for the GP training in Utrecht or elsewhere.

After the first three months of training one-fourth of the trainees had received a score 'poor' or 'insufficient' on one or more competency roles and 15 % a score 'insufficient' on the LHK. There was a moderate correlation between the three competency roles (rho: 0.43 – 0.58), but no correlation between the competency roles and LHK (rho: 0.06 – 0.15). In the fourth quarter, most trainees scored 'sufficient' on the three roles: 98.5 % as 'medical expert' , 95.4 % as 'communicator', 91.9 % as 'professional' and 85 % scored 'sufficient' on the LHK. 194 (90.2 %) trainees were directly admitted to the second year (Figure 2). Twenty-one trainees received a 'go, unless' (8.4 %) or 'no-go' decision (1.4 %). Half of them (*n* = 10) showed shortcomings in one role with an equal distribution ('medical expert' *n* = 3; 'communicator' *n* = 4; 'professional' *n* = 3). The others (*n* = 11) had problems in two or three roles, with an equal distribution across the three roles. Half of the identified shortcomings were reflected in an 'insufficient' competency assessment score. Two trainees showed shortcomings in another role ('manager', 'scholar').

Older trainees and trainees with score 'insufficient' on competency role 'medical expert' or LHK in the first quarter were more likely to exhibit poor performance and involuntary attrition (Table 2); adj OR 1.1 (CI 1.0 – 1.3), adj OR 8.9 (CI 3.0 – 26.33), adj OR 2.1 (CI 1.1 – 4.0) respectively.

Discussion

Twenty-one of 215 trainees exhibited poor performance in their first year. Ultimately, three trainees dropped out involuntary. Trainees with score 'insufficient' on competency role as 'medical expert' and on medical GP knowledge at the beginning of the program are more likely to exhibit poor performance at the end of year one, just like older trainees. In the literature, attrition rates of trainees of internal medicine or surgery are mentioned divergent and higher.[10-12]

In this analysis we restricted to three competency roles - 'medical expert', 'communicator' and 'professional' - as these roles are the focus in the first year.[4] The trainees who exhibited poor performance or dropped out had shortcomings in these competency roles; two trainees also had a problem in another role. In the literature the same roles are mentioned to reveal shortcomings.[10-14]

In only half of the cases the shortcomings of the trainees were reflected in an insufficient competency score. This is a known phenomenon. Trainers and tutors identify shortcomings and insufficient progress, but they do not always express these in quantitative assessments.[15-16] This finding supports the argument of van der Vleuten et al. to combine summative and formative assessments in medical education.[17] Further assessment training with attention to attitude and motivation is required for trainers and staff members.[18]

At the start of the training there was a moderate correlation between the competency scores on the three domains, but no correlation with the LHK scores. A possible explanation might be that competencies are on another level of Millers' pyramid than knowledge.[19] The LHK takes place on 'knows' and 'knows how-level' while the competency assessments are localized at the 'does'-level. In the literature, a correlation between knowledge tests and competency ratings is reported.[20-23] We expect that correlation will increase during the training.

Age was independently associated with poor performance/involuntary attrition. The OR of 1.1 (CI 1.0 – 1.3) means that the risk to exhibit poor performance or drop out increases per year with 10%. In postgraduate training of surgery and gynaecology in the USA higher age was also associated with attrition.[24-25] As underlying reason especially living conditions, such as having family responsibilities were mentioned. Maybe already developed professional behaviours in older trainees might play a role. More or less clinical experience was not associated with poor performance/involuntary attrition. This result is consistent with earlier findings that the relationship between the amount of clinical experience and performance of the trainee is complex.[26-27] The fact that diverging clinical experience was put together and was just quantitatively measured could play a role.

Limitations of this study

In this study, we used routinely gathered data. No additional information was collected on other possible determinants such as personal and health problems. The analysis focuses on the first year as this is a crucial phase of the training and almost half of the total drop outs takes place in this year.[2,7]

The fact that within the cohort, trainees were assessed on two different checklists (LEV list and Compass), we solved by using standardized values (z-scores).

We assume that the internal validity, despite the relatively small number of trainees with poor performance/involuntary attrition is satisfactory thanks to imputation of missing values.[8-9] Because the preferred training areas is particularly geographically based, there is no indication that the results would not apply to other Dutch postgraduate GP trainings.[28] To what extent the results are valid to other postgraduate training programs in the Netherlands with similar competency assessment lists has to be investigated.

Conclusion

This study shows that 'high-risk' trainees can be identified early in the training. We recommend to monitor this group, just like the elder trainees, and provide additional guidance with more attention to Individual Development Plans and own responsibility in the learning process.

The importance of this study lies in the description of the actual course of the GP training. This cohort will be followed during the rest of the GP training in order to investigate overall poor performance and attrition. We advise to set up a systematic national registration system to identify determinants of stagnation. In addition, the results of this study raise the question if knowledge and competencies, conform the English GP training should be involved in the selection procedure of GP trainees.[29-30]

References

1. Katzenbauer M. 'That culture did not suit me'. (In Dutch: 'Die cultuur paste niet bij mij'). Med Cont 2009;64:580-3.
2. Kuyvenhoven MM, Vermeulen MI, van Campen SM, Schmidt JE. Many trainees drop out. (In Dutch: Veel aiossen haken af). Med Cont 2010;65:2206-7.
3. Vermeulen MI, Rijksen M, Pieters HM, Kuyvenhoven MM. Trainees during their first year of the GP training. (In Dutch: Artsen in opleiding tijdens het eerste jaar van de Huisartsopleiding). Huisarts Wet 2009;52:349.
4. Framework. Baarveld F, Bottema BJAM, Bueving HJ, Langendoen-Roel M, van Leeuwen YD ea. Utrecht, SVUH 2005 (SVUH0502).
5. Competence profile of the GP. (In Dutch: Competentieprofiel van de huisarts) Berkestijn van LGM, Duin van BJ, Hoekstra M, Maiburg M, Oosterling EMP ea. Utrecht, PVH 2005.
6. van Leeuwen YD Pollemans MC, Mol SS, Eekhof JA, Grol R, Drop MJ. The Dutch knowledge test for general practice: issues of validity. Eur J Gen Pract 1995;1:113-7.
7. Framework Decision. College of General Practice and Nursing Home Medicine (In Dutch Kaderbesluit CHVG. College voor Huisarts en Verpleeghuis Geneeskunde); Staatscourant, 25 november 2004, nr. 228.
8. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.
9. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087-91.
10. Yao DC, Wright SM. The challenge of problem residents. J Gen Intern Med. 2001;16:486-92.
11. Williams RG, Roberts NK, Schwind CJ, Dunnington GL. The nature of general surgery resident performance problems. Surgery 2009;145:651-8.
12. Bergen PC, Littlefield JH, O'Keefe GE, Rege RV, Antony TA, Kim LT ea. Identification of high risk residents. J Surg Res 2000;92:239-44.
13. Longo WE, Seashore J, Duffy A, Udelsman R. Attrition of categoric general surgery residents: results of a 20-year audit. Am J Surg 2009;197:774-8.
14. Kohanzadeh S, Hayase Y, Lefor MK, Nagata Y, Lefor AT. Factors affecting attrition in graduate surgical education. Am Surg 2007;73:963-6.
15. Williams RG, Dunnigton GL, Klamen DL. Forecasting residents' performance - partly cloudy. Acad Med 2005;80:415-22.
16. Dudek NL, Marks MB, Regehr G. Failure to fail: The perspectives of clinical supervisors. Acad Med 2005;80:S84-S87.
17. van der Vleuten CP, Schuwirth LW, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. Best Pract Res Clin Obstet Gynaecol 2010;24:703-19.
18. Govaerts MJ, van der Vleuten CP, Schuwirth LW, Muijtjens AM. Broadening perspectives on clinical performance assessment: Rethinking the nature of In-training assessment. Adv Health Sci Educ 2007;12:239-60.
19. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65:S63-S67.
20. Thundiyil JG, Modica RF, Silvestri S, Papa L. Do United States medical licensing examination (USMLE) scores predict in-training test performance for emergency medicine residents? J Emerg Med 2010;38:65-9.
21. Shellito JL, Osland JS, Helmer SD, Chang FC. American Board of Surgery: can we identify surgery residency applicants and residents who will pass the examinations on the first attempt? Am J Surg 2010;199:216-22.
22. Spitzer AB, Gage MJ, Looze CA, Walsh M, Zuckerman JD, Egol KA. Factors associated with successful performance in an orthopaedic surgery residency. J Bone Joint Surg Am 2009;91:2750-5.
23. Perez JA Jr, Greer S. Correlation of United States Medical Licensing Examination and Internal In-Training Examination performance. Adv Health Sci Educ 2009;14:753-8.
24. Naylor RA, Reisch JS, Valentine RJ. Factors related to attrition in surgery residency based on application data. Arch Surg 2008;143:647-51.

25. McAlister RP, Andriole DA, Brotherton SE, Jeffe DB. Attrition in residents entering US obstetrics and gynecology residencies: analysis of National GME Census data. Am J Obstet Gynecol 2008;199: 574.e1-6.
26. Wimmers PF, Schmidt HG, Splinter TA. Influence on clerkship experiences on clinical competence. Med Educ 2006;40:450-8.
27. Martin IG, Stark P, Jolly B. Benefiting from clinical experience: the influence of learning style and clinical experience on performance in an undergraduate objective structured clinical examination. Med Educ 2000;34:530-4.
28. Unpublished survey: “What are reasons for trainees to choose a GP training department”? (In Dutch: “Redenen voor de keuze waar aios hun opleiding tot huisarts volgen.”) Tromp F, Mokkink HGA, Thoonen BPA, Bottema BJAM. VOHA Nijmegen 2009.
29. Patterson F, Ferguson E, Norfolk T, Lane P. A new selection system to recruit general practice registrars: preliminary findings from a validation study. BMJ 2005;330:711-4.
30. Patterson F, Baron H, Carr V, Plint S, Lane P. Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. Med Educ 2009;43:50-7.

Table 1 Individual characteristics, early competency and knowledge assessments of the trainees (*n* = 215)

Individual characteristics	
Gender, % male	31.2
Age in years; mean (SD)	29.5 (4.2)
Region medical school, % NW Europe	93.1
Europe Past performance, % < 1 year as MD	48.3
Times of application, % first	76.7
Early competency and knowledge assessments, % 'sufficient, good'	
'Medical expert '	74
'Communicator'	71
'Professional'	78
National GP Knowledge Test	85

Table 2 Association between age (continuous), assessment score on 'medical expertise', knowledge and poor performance/involuntary attrition

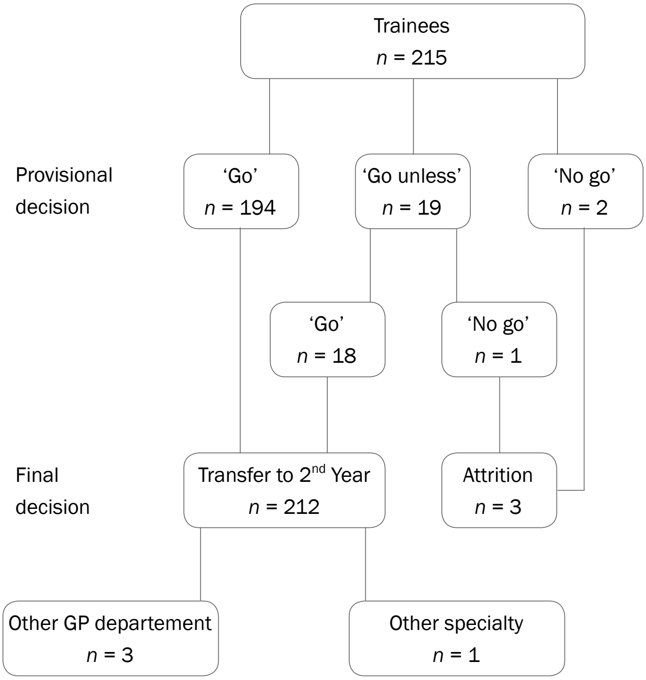
Individual characteristics	Multivariate OR's	95% Confidence interval
Age	1.1	1.0 – 1.3
Early assessment scores		
Medical expert	2.1	1.1 – 4.0
Knowledge (0: sufficient; 1:insufficient)	8.9	3.0 – 26.3

Figure 1 Example of items and assessment scale from the *Compass*

Indicators	Competencies
Acting in context • Mentions contextual factors noted in the medical file that may be related to the complaint if necessary	*1.1* Interprets symptoms in context
Diagnostic skills • Demonstrates knowledge and understanding of diagnostic and therapeutic arsenal of the GP	□ □ □ □ *1st 2nd 3rd 4th* *Quarter*
Evidence based medicine • Uses the Standards of the Dutch Royal College of GPs or other evidence based guidelines and recommendations in the scientific research appropriately • Provides rational substantiation for decisions towards diagnostics and policies based on epidemiologic data, evidence-based guidelines and (reflection on) experience	*1.2* Applies the diagnostic and therapeutic arsenal of the profession in an appropriately and evidence based way □ □ □ □
Logical structuring of the contact • Masters the complete spectrum of problem identification, history taking, physical and additional examination, providing information, advice, counselling and referral • Provides care appropriately by adopting a logical approach: information gathering, making (provisional) diagnosis and deciding on treatment or policy	*1.3* Provides primary care in a systematic way □ □ □ □
Overall score for domain medical expertise The medical expertise of the GP includes all medical activities he/she engages in response to complaints, problems and questions about health and disease. The essence of medical actions involves identifying the nature and seriousness of the complaint and assessing the need for intervention. From a working hypothesis, a treatment plan is generated, and the effect is monitored.	□ □ □ □

Rating scale:
4 = excellent (maintain this standard)
3 = satisfactory (keep working on it)
2 = unsatisfactory (concentrate on this specifically)
1 = very weak (requires urgent attention)
? = unclear (lack of information)

Figure 2 Flowchart of the trainees from 1st to 2nd year of the training



Chapter 5

Risk factors for poor performance among trainees in a Dutch postgraduate GP training program

Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, Pieters HM, van der Graaf Y, Damoiseaux RAMJ

Abstract

Introduction

The early detection of poor performance among trainees in a postgraduate medical program might be fruitful because it could enhance the success of remediation efforts.

Objectives

To explore the frequency, nature and risk factors of poor performance among Dutch postgraduate GP trainees.

Methods

All trainees who started the program between 2005 and 2007 were included in the study cohort. The following data were gathered from clerical dossiers: individual characteristics; early assessments of trainee knowledge and competency roles of ‘medical expert’, ‘communicator’ and ‘professional’; and training process characteristics (e.g., illness during the previous year). The outcome measured was poor performance that occurred once or more during training, as formally assessed by trainers and staff members. Associations between individual characteristics, early assessment scores and outcome were measured using multivariate logistic regression analysis. Additionally, sub analyses were performed for each year.

Results

Two hundred fifteen trainees started GP training, and 49 exhibited poor performance (9.7%, 13.0% and 7.4% in training years 1, 2 and 3, respectively; 22.8% overall). Six trainees (2.8%) were dismissed. In the 1st and 2nd years, problem areas among poor performers were distributed equally across the roles of ‘medical expert’, ‘communicator’ and ‘professional’. In the 3rd year, a shortcoming in ‘professionalism’ was the most common problem. Higher age was a risk factor for poor performance; OR 1.16 (CI 1.06 – 1.27). Trainees with sufficient assessment scores in ‘communication’ and knowledge were at lower risk of poor performance; OR 0.50 (CI 0.33 – 0.77) and OR 0.16 (0.07 – 0.40), respectively. Poor performance in the previous year was a risk factor for poor performance in the 2nd and 3rd years; OR 4.20 (CI 1.31 – 13.47) and OR 5.40; (CI 1.58 – 18.47), respectively.

Conclusions

Poor performance is prevalent, primarily occurring within a single training year. This finding suggests that trainees and their trainers are capable of solving trainee problems. Higher age, insufficient assessment scores early in the training and poor performance in a previous year constitute risk factors for poor performance. In light of these findings, we recommend additional monitoring of trainees who are ‘at risk’ in these respects.

Introduction

Attrition (both voluntary and involuntary) is a common problem in postgraduate medical training programs, with rates varying depending on specialty. In the United States, attrition rates range from 3.3% in psychiatry to 8.6% in family medicine and almost 30% in some surgery programs.[1-8] In the Netherlands, approximately ten per cent of internal medicine and surgery trainees withdraw from their programs prematurely.[9] From 2002 to 2007, the overall attrition rate among all Dutch GP training programs was 7.5 %, and the percentage of involuntary attrition (i.e., attrition due to incompetence) was 1.9%.[10] While trainers and staff members are known to be reluctant to confront trainees with their shortcomings and, if necessary, impose consequences, the involuntary attrition rate might underreport the number of trainees with shortcomings in competencies — that is, trainees exhibiting poor performance.[11-13-15]

Trainees who exhibit poor performance have significant problems with knowledge, attitudes or skills that require monitoring and remediation.[14] These issues often create a significant burden for the trainee, trainers and staff members.[16-18] Trainee shortcomings can be classified effectively using the CanMEDS framework.[19] Early detection of poor performance could be fruitful, as it would provide remediation efforts with a greater chance to succeed.[4,12,14,17] However, some performance deficiencies, particularly those related to attitudinal problems, can be resistant to remediation and remain 'chronic'.[11,18,20]

In light of these considerations, we explored the extent and nature of poor performance in a Dutch postgraduate GP training program in order to identify possible risk factors of 'trainees at risk' early in training. The following research questions were posed:

- How many trainees exhibit poor performance during the three-year program, and in which competency roles do shortcomings appear?
- To what degree are individual characteristics (e.g., gender, age and selection scores) and early competency/knowledge assessments associated with poor performance?
- To what degree are individual characteristics, early competency/knowledge assessments and training process characteristics (e.g., illness or performance in the previous year) associated with poor performance in each year?

Methods

The Utrecht Trainee Competence Study in General Practice (GP-UTCS) is a retrospective observational cohort study of all trainees who started the GP training in Utrecht between March 1st, 2005 and September 1st, 2007. At the time of the study, ethical approval was not mandatory for routinely gathered data. We executed the study according to 'the code of conduct' for the use of personal data in scientific research.[21] Prior to data processing, all data were clerically anonymised.

Postgraduate GP training

The Dutch postgraduate GP training takes three years. The 1st and 3rd years provide practical training in general practice under the supervision of an experienced GP (the 'trainer'). The 2nd year is dedicated to hospital rotations, primarily in the emergency room (six months), nursing homes (three months) and mental health institutions (three months). Dispensation is offered for one or more rotations during the 2nd year in cases where trainees have received relevant previous experience in authorised institutions. Each week, the trainee attends a one-day tutorial in a small group of 12 trainees, with two staff members (a general practitioner and a psychologist) providing training in theoretical, practical and reflective skills. Every three months, each trainee's performance is evaluated by his or her GP trainer or intramural supervisor and staff members. Twice a year, the trainees' knowledge progress is assessed via a National GP Knowledge Test.[22] At the end of each year, relying on all assessments, the head of the training department decides whether a trainee can continue with the program ('go') or whether certain conditions must be met before he/she is allowed to continue ('go, unless'). These conditions are individualised requirements designed to help trainees improve in particular problem areas; they might include additional time in the program, further competency or knowledge training and assessment or an extra rotation in another clinical setting. If a trainee does not meet these conditions, he/she is dismissed ('no go').

Data collection

All data were derived from trainee dossiers.

Individual characteristics. The trainee characteristics used for this analysis included gender; age (in years) at the time of entry into the program; the region where medical school was completed (NW Europe; elsewhere); past performance - clinical experience as a medical doctor (< one year or ≥ one year); the number of times the trainee applied to the training program (once; more than once); the duration of dispensation (0, 3, 6 or ≥ 9 months) and his/her mean selection score on a three-point scale - i.e., below standard (1), standard (2) and above standard (3), as evaluated by a selection committee member (a staff member, trainer, and trainee) with respect to the following domains: motivation, orientation on the job, learning needs and personal attributes.[23]

Early assessments. After the first three months of training, trainees were evaluated in three competency roles ('medical expert', 'communicator' and 'professional') using a four-point scale (1 = poor, 2 = insufficient, 3 = sufficient and 4 = good). Additionally, in the second month of the program, trainees completed the National GP Knowledge Test, where their performance was scored as either insufficient or sufficient.[22]

Training process characteristics. These characteristics are time-dependent variables that may vary from year to year during training. These characteristics include the percentage of time the trainee spent on weekly employment (< 85%, 85-95%, > 95%); the duration of any recorded illnesses (in weeks) that lasted more than two weeks; the frequency of maternity leaves and the quality of the trainee's performance during the previous year (as assessed via the categorisations of 'go', 'go, unless', or 'no go').

Outcome measure. The outcome measure was poor performance ('go, unless' or 'no go'), as assessed by the head of the department at the end of each year based on the assessments of the trainers and staff members. To determine the shortcomings of the poor performers, we analysed trainee dossiers. Using the CanMEDS framework, we defined trainee weaknesses using the following roles: 'medical expert' (including medical knowledge), 'communicator', 'collaborator', 'manager', 'health advocate', 'scholar' and 'professional' (which encompasses showing respect to others, self-care, integrity and reflection).[24]

Analysis

First, trainees' individual characteristics and early competency/knowledge assessment scores were collected. Next, the course of the cohort was determined, and training process characteristics were provided. The associations between individual characteristics, early assessments and outcomes were estimated using univariate and multivariate logistic regression analysis (odds ratios; 95% confidence interval) as follows: first, the association between individual characteristics and outcome was measured and then the association between early assessments and outcome. Variables with a p-value ≤ 0.10 after backward analysis, age and gender were taken into account in the final backward model. Using a similar approach, sub analyses of the data for each year were performed. In these models, the training process characteristic 'percentage weekly employment' was added. In the models for years 2 and 3, all training process characteristics for the previous year(s) were included. There were some missing data (representing 1.3% of all data). To address this issue, we applied multiple imputation, as incomplete case analysis (excluding trainees with one or more missing values from the analysis) could lead to a loss of statistical power and biased results.[25] Analyses were performed using SPSS version 20.

Results

Individual characteristics. In total, 215 trainees started GP training between March 1st, 2005, and September 1st, 2007 (Table 1). One-third of these trainees were male, and the mean age was 29.5 years (SD 4.2). Most trainees (93.0%) had completed medical school in north western Europe. Almost half of them had worked less than one year as a medical doctor at the time of application, and one-fourth had applied for the GP training program two or more times. The mean selection score was 2.4 (SD 0.3). More than one-fourth had no dispensation, 13% had three months, 43% had six months and the remaining trainees had nine or twelve months.

Early assessments. After the first three months of training, approximately one-fourth of the trainees had received a score of 'poor' or 'insufficient' with respect to the roles as 'medical expert', 'communicator' or 'professional'. Fifteen per cent received a score of 'insufficient' on the GP knowledge test.

Training process characteristics. The percentage of trainees who maintained full-time employment (> 95%) decreased from about 80% in the 1st year to 70% in the 2nd year and 60% in the 3rd year (Table 2). Sixteen per cent had been ill for more than two consecutive weeks. Among these individuals, the mean duration of illness was 10.6 (SD 8.4) weeks (median 7.5 weeks). During the course of the program, of the 148 female trainees, 47 went on maternity leave once and 19 twice. Most maternity leaves occurred in the 3rd year (when there were 42, as compared with 18 in the 1st year and 25 in the 2nd year).

Poor performance. In the 1st year, 21 of the 215 trainees (9.7%) were assessed as exhibiting poor performance, and three of them were eventually dismissed (Figure). Four trainees voluntarily left the program after year 1; three of them went to another GP program, and one changed specialisation. One trainee received dispensation for all rotations; thus, 207 trainees entered year 2. In that year, twenty-seven trainees (13%) exhibited poor performance. Of these, nineteen had not experienced problems in the 1st year (incidence: 9.2%). Eight of the 18 trainees who exhibited poor performance in year 1 (44.4%) did so in year 2 as well, whereas ten did not. At the end of year 2, one trainee was dismissed and four trainees left for another GP department. Eventually, 203 trainees entered year 3. Fifteen trainees exhibited poor performance in that year. Nine trainees exhibited problems for the first time (4.4%), two trainees for the second time and four for the third time. Two trainees were dismissed in the last year of the program, and one trainee left the program for another GP department. The involuntary attrition rate of this cohort was 2.8%; among trainees who exhibited poor performance, the rate was 12.2%. One additional trainee left voluntarily to go to another postgraduate program; thus, the total attrition rate was 3.3%.

Overall, 49 trainees (22.8%) exhibited poor performance: 39 in one of the three years, six in two different years and four in every year of the program. Trainees exhibiting poor performance primarily demonstrated shortcomings in the following three roles: 'medical expert', 'communicator' and 'professional' (Table 3). In years 1 and 2, trainee shortcomings were split equally among these three roles; in year 3, shortcomings were primarily related to the role of 'professional'. The number of shortcomings decreased the longer the poor performer was in training. In year 1, more than half of these trainees had shortcomings in two or more competencies; in year 2, the proportion was one-third and, in year 3, only one-fourth of the trainees had shortcomings in two or more competencies.

Risk factors for poor performance. Older trainees ran a greater risk of exhibiting poor performance during training (Table 4); OR 1.16 (CI 1.06 – 1.27), which means that the risk of showing poor performance increases 16% per year. Trainees with sufficient scores in 'communication' and knowledge after three months in training were at lower risk of exhibiting poor performance; OR 0.50 (CI 0.33 – 0.77) and OR 0.16 (CI 0.07 – 0.40), respectively. Sub analyses for each training year showed similar results (Table 6); higher age was a risk factor for poor performance in the 1st and 2nd years; OR 1.14 (CI 1.02 – 1.26) and OR 1.12 (CI 1.02 – 1.23), respectively; insufficient scores in 'medical expertise' and knowledge in the 1st year; OR 0.49 (CI 0.29 – 0.84) and OR 0.12 (CI 0.04 – 0.35), respectively or 'communication' in the 2nd year; OR 0.47 (CI 0.30 – 0.75). Poor performance in the 1st and 2nd years was independently associated with poor performance in years 2 and 3; OR: 4.20 (1.31 – 13.47) and OR: 5.40; (CI 1.58 – 18.47), respectively. Further analysis of gender and extreme poors (i.e., individuals who were dismissed or exhibited poor performance in two or three years; *n* = 14) revealed comparable findings (not in table).

Discussion

Main findings

Forty-nine trainees exhibited poor performance, most of them in single year, and ten exhibited poor performance in two or three years. Six trainees were eventually dismissed. Most trainee shortcomings fell into the roles of 'medical expert', 'communicator' and 'professional'. In years 1 and 2, trainee weaknesses were distributed equally across these three competency roles, whereas in year 3, shortcomings primarily pertained to trainee 'professionalism'. Higher age and insufficient scores in 'communication' and knowledge were risk factors for poor performance. In the 1st and 2nd years, age was a risk factor, just as insufficient scores in certain roles (i.e., 'medical expert' and knowledge in the 1st year and 'communicator' in the 2nd). Poor performance in the previous year was a risk factor for poor performance in years 2 and 3.

Strengths and limitations

A cohort of 215 trainees was followed throughout their postgraduate GP training. As far as we know, this is the first study that has tracked GP trainees over consecutive years. We are aware that many determinants regarding the number of outcomes have been used in the model, but our results were stable, and the overall cohort results were corroborated by results for each training year. This study was performed using clerically collected data, which is standard operating procedure. Use of clerical data resulted in a limitation, however: the study design did not allow us to collect data regarding additional possible risk factors, such as personality traits, medical school records or the nature of any recorded illnesses.[26,27]

We assume that the study's internal validity is satisfactory, despite the relatively small number of trainees with poor performance, due to imputation of missing values.[25] It is unclear whether our findings are generalisable to other Dutch GP training departments. However, we can reasonably assume that our trainees and the program's trainee monitoring system, which is based on national assessment regulations, do not differ from those of other Dutch postgraduate GP training programs. By exploring the dossiers of the poor performers, we have attempted to choose the most valid source of information that offers the least risk of bias with respect to trainee shortcomings.

Discussion of main findings

Twenty-three per cent of all trainees exhibited poor performance in one or more years during their postgraduate GP training program, meaning that these trainees had significant problems with knowledge, attitude or skills.[14] Corresponding figures from other Dutch postgraduate training programs have not been obtained. A comprehensive range of attrition rates across medical disciplines have been mentioned; there are low rates (<6%) for psychiatry and geriatric medicine, median rates for internal and family medicine (8-15%) and higher rates (20-30%) for surgical training programs.[4,5,11,18,28,29] In Canada, the mean rate of 'residents in difficulty' across several postgraduate training programs was 3%.[19] However, it is difficult to compare these percentages because they are influenced by documentation, identification criteria and the assessment culture of the department.

Almost five per cent of trainees exhibited poor performance in two or three of their training years, and their shortcomings mostly remained in the same competency roles. This finding may indicate that remediation efforts were ineffective, trainee problems were resistant to these efforts, or decisions regarding dismissal (which are often difficult) were postponed.[13] We suppose that these trainees could be at risk of exhibiting poor performance as future GPs.[18,30-33]

In years 1 and 2, approximately ten per cent of trainees exhibited poor performance, whereas in year 3, this percentage was lower. Reamy also found that most trainee problems were identified during the first two years of a family medicine program.[20] The changes in setting that occur during training may contribute to this asymmetry. In year 1, trainees are entering a new GP practice, and in year 2, they are starting rotations; in year 3, however, they return to a GP practice, a familiar educational environment.

Among the trainees in this study, poor performance primarily pertained to shortcomings as a 'medical expert', 'communicator' and 'professional.' This finding aligns with previous research that cited insufficient knowledge, poor clinical judgement, poor communication, poor interpersonal skills and attitudinal problems as reasons for poor performance.[4,11,14,17,18,20,28] Almost half of the poor performers had shortcomings in two or more competency roles, findings that echo other studies.[4,11,17,18] The distribution pattern of trainee shortcomings changed over time. During the first two years, trainee shortcomings were distributed equally across the roles as 'medical expert', 'communicator' and 'professional'; in the 3rd year, however, most shortcomings fell into the role as 'professional'. If a trainee exhibited poor performance in two or three years, the role 'professionalism' was always involved. This finding suggests that shortcomings as a professional could be more resistant to remediation.[11,18,20]

Risk factors

Higher age is a risk factor for poor performance in years 1 and 2. Some of the previous literature on this topic mentioned higher age as a risk factor for poor performance and attrition, while other studies did not support this relationship.[3,4,7,32,34,35] It may be that older trainees have more family responsibilities or possess out-dated knowledge, or they may be less resilient or less flexible in adapting to new clinical situations or academic settings. Early insufficient assessments in the roles as 'medical expert' or 'communicator' and knowledge were associated with poor performance among trainees in the first two years. The link between earlier and later performance within the same training or school mirrors findings of others.[4,16,36-39] This finding suggests the utility of performing early competency and knowledge assessments in a programmatic assessment approach.[28,40-44] Early insufficient assessments in the role of 'professional' did not constitute an independent risk factor, probably because this domain is so broad that it is hard to assess early in the GP training. Poor performance in a previous year was a strong risk factor for poor performance in the next one, which corroborates earlier findings.[3,4] These results may indicate that poor performance is hard to remedy and, thus, can become 'chronic'.

In our study, mean application score was not associated with poor performance. This finding aligns with earlier studies that indicate it is difficult to predict performance problems from application data.[16,45] Likewise, illness was not associated with poor performance, even though health problems, substance abuse and psychiatric illness are often mentioned in the literature as possible reasons for voluntary and involuntary attrition.[4,19,28,32] Part-time employment and maternity leave were not associated with poor performance; only programmatic schedulers appear to be disadvantaged by the special circumstances that they require.

Conclusion

Poor performance among trainees in a postgraduate GP training program is quite common. Most trainees exhibited poor performance in only one year of the program under examination, suggesting that trainees and their trainers are capable of solving trainee problems. However, higher age, insufficient early assessments and poor performance in a previous year represent risk factors for poor performance during postgraduate GP training. Early assessments might be useful in addressing these risk factors. Because they are at risk of repeated or on-going problems, poor performers need extra monitoring and supervision by the trainers and staff.

References

1. Roback HB, Crowder MK. Psychiatry resident dismissal. A national survey of training programs. Am J Psychiatry 1989;146:96-8.
2. Laufenburg HF, Turkal NW, Baumgardner DJ. Resident attrition form family practice residencies: United States versus international medical graduates. Fam Med 1994;26:614-7.
3. Yeo H, Bucholz E, Ann Sosa J, Curry L, Lewis FR, Jones AT, Viola K, Lin Z, Bell RH. A national study of attrition in general Surgery training : which residents leave and where do they go? Ann Surg 2010;252:529-34.
4. Bergen PC, Littlefield JH, O'Keefe GE, Rege RV, Antony TA, Kim LT ea. Identification of high risk residents. J Surg Res. 2000;92:239-44.
5. Yaghoubian A, Galante J, Kaji A, Reeves M, Melcher M, Salim A, Dolich M, de Virgilio C. General Surgery Resident remediation and attrition. Arch Surg 2012;147:829-33.
6. Longo WE, Seashore J, Duffy A, Udelsman R. Attrition of categoric general surgery residents: results of a 20-year audit. Am J Surg. 2009;197:774-8.
7. Sullivan MC, Yeo H, Roman SA, Ciarleglio MM, Cong X, Bell RH jr, Ann Sosa J. Surgical residency and attrition: defining the individual and programmatic factors predictive of trainee losses. J Am Coll Surg 2013;216:461-71.
8. Dodson TF, Webb ALB Why do residents leave general surgery? The hidden problem in today's program. Curr Surg 2005;62:128-31.
9. Katzenbauer M. 'That culture did not suit me'. (In Dutch: 'Die cultuur paste niet bij mij'). Med Cont 2009;64:580-3.
10. Kuyvenhoven MM, Vermeulen MI, van Campen SM, Schmidt JET. Many trainees drop out. (In Dutch: veel aiossen haken af.) Med Cont. 2010;65:2206-7.
11. Williams RG, Roberts NK, Schwind CJ, Dunnington GL. The nature of general surgery resident performance problems. Surgery 2009;145:651-8.
12. Evans DE, Alstead EM, Brown J. Applying your clinical skills to students and trainees in academic difficulty. Clin Teach 2010;7:230-5.
13. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. Acad Med 2005; 80 (10 suppl):S84-S87.
14. Steinert Y. The "problem" learner: Whose problem is it? AMEE Guide No. 76. Med Teach. 2013;35:e1035-e1045.
15. Adams KE, Emmons S, Romm J. How resident unprofessional behavior is identified an managed: a program director survey. Am J Obstet Gynecol 2008;198:692.e1-692.e5.
16. Brenner AM, Mathai S, Jain S, Mohl PC. Can we predict "problem residents"? Acad Med 2010;85:1147-51.
17. Yao DC, Wright SM. The challenge of problem residents. J Gen Intern Med. 2001;16:486-92.
18. Williams RG, Dunnigton GL, Klamen DL. Forecasting residents' performance- Partly clouded. Acad Med. 2005;80:415-22.
19. Zbieranowski I, Takahashi SG, Verma S, Spadafora SM. Remediation of residents in difficulty: a retrospective 10-year review of the experience of a postgraduate board of examiners. Acad Med 2013;88:1-6.
20. Reamy BV, Harman JH. Residents in trouble: An in-depth assessment of the 25- year experience of a single family medicine residency. Fam Med 2006;38:252-7.
21. Association of Universities the Netherlands. Code of conduct for use of personal data in scientific research (http://www.vsnu.nl/code-pers-gegevens.html, in Dutch) 2005.
22. Van Leeuwen YD, Pollemans MC, Mol SSL, Eekhof JAH, Grol R, Drop MJ. Dutch knowledge test for general practice: issues of validity. Eur J Gen Pract 1995;1:113-7.
23. Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, Graaf van der Y, Damoiseaux RAMJ. Dutch postgraduate GP selection procedure; reliability of interview assessments. BMC Family Practice 2013,14:43 doi:10.1186/1471-2296-14-43.
24. Frank JR. The CanMEDS 2005 Physician Competency Framework. Better standards. Better physicians. Better care. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2005.
25. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: A gentle introduction to imputation of missing values. J of Clin Epidemiol. 2006;59:1087-91.

26. Lievens F, Coetsier P, De Fruyt F, De Maeseneer J. Medical student's personality characteristics and academic performance: a five factor model perspective. Med Educ 2002;36:1050-6.
27. Hamdy H, Prasad K, Anderson MB, Scherpbier A, Williams R, Zwierstra R, Cuddihy H. BEME systematic review: predictive values of measurements obtained in medical schools and future performance in medical practice. Med Teach. 2006;28:103-16.
28. Mitchell C, Bhat S, Herbert A, Baker P. Workplace based assessments of junior doctors: do scores predict training difficulties? Med Educ 2011;45:1190-8.
29. ABIM: "The problem resident." VHS videocassette produced by American Board of Internal medicine; Portland, Ore 1992.
30. Papadakis MA, Hodgson CS, Teherani A, Kothatsu ND. Unprofessional behaviour in medical school is associated with subsequent disciplinary action by a state medical board. Acad Med 2004:79:244-9.
31. Resnick AS, Mullen JL, Kaiser LR, Morris JB. Patterns and predictions of resident misbehaviour- a 10-year retrospective look. Curr Surg 2006;63:418-25.
32. Yao DC, Wright SM. National survey of internal medicine residency program directors regarding problem residents. JAMA 2000;284:1099-1104.
33. Tamblyn R1, Abrahamowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D, Smee S, Blackmore D, Winslade N, Girard N, Du Berger R, Bartman I, Buckeridge DL, Hanley JA Physician scores on national clinical skills examination as predictors of complaints to medical regulatory authorities. JAMA. 2007;298:993-1001.
34. Naylor RA, Reisch JS, Valentine RJ. Factors related to attrition in surgery residency based on application data. Arch Surg. 2008;143:647-51.
35. McAlister RP, Andriole DA, Brotherton SE, Jeffe DB. Attrition in residents entering US obstetrics and gynecology residencies: analysis of National GME Census data. Am J Obstet Gynecol. 2008;199:574.e1-6.
36. Van Leeuwen YD, Mol SS, Polleman MC van der Vleuten CP, Grol R, Drop MJ. Selection for postgraduate training for General Practice: role of knowledge tests. Br J Gen Pract 1997;47:359-62.
37. Winston KA, van der Vleuten CPM, Scherpbier AJJA. Prediction and prevention of failure: an early intervention to assist at-risk medical students. Med Teach 2014:36:25-31.
38. Yates J. Development of a 'toolkit' to identify medical students at risk of failure to thrive on the course: An exploratory retrospective case study. BMC Med Educ 2011 11:95. doi:10.1186/1472-6920-11-95
39. Stegers-Jager KM, Cohen-Schotanus J, Themmen APN. 2013. The effect of a short integrated study skills programme for first-year medical students at risk of failure: A randomised controlled trial. Med Teach 2013;35:120–126.
40. Van der Vleuten CPM, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005;39:309-17.
41. Thundiyil JG, Modica RF, Silvestri S, Papa L. Do United States medical licensing examination (USMLE) scores predict in-training test performance for emergency medicine residents? J Emerg Med. 2010;38:65-9.
42. Shellito JL, Osland JS, Helmer SD, Chang FC. American Board of Surgery: can we identify surgery residency applicants and residents who will pass the examinations on the first attempt? Am J Surg. 2010;199:216-22.
43. Spitzer AB, Gage MJ, Looze CA, Walsh M, Zuckerman JD, Egol KA. Factors associated with successful performance in an orthopaedic surgery residency. J Bone Joint Surg Am.2009; 91: 2750-5.
44. Perez JA Jr, Greer S. Correlation of United States Medical Licensing Examination and Internal In-Training Examination performance. Adv Health Sci Educ. 2009;14:753-8.
45. Dubovsky SL, Gendel M, Dubovsky AN, Rosse J, Levin R, House R. Do data obtained from admission interview and resident evaluations predict later personal and practice problem. Acad Psych 2005;29:443-7.

Table 1 Individual characteristics, early competency and knowledge assessments of the trainees (*n* = 215)

Individual characteristics	
Gender, % male	31.2
Age in years; mean (SD)	29.5 (4.2)
Region medical school, % NW Europe	93.1
Past performance, % < 1 year as MD	48.3
Times of application, % first	76.7
Mean selection score: mean (SD)	2.4 (0.3)
Dispensation, %	
0 month	27.9
3 months	12.6
6 months	43.3
9 months	13.5
12 months	2.8 (*n* = 6*)
Early competency and knowledge assessments, % 'sufficient, good'	
Medical expert	73.5
Communicator	71.3
Professional	77.5
GP knowledge test	84.6

**n = 6: 1 trainee received 12 month dispensation and could directly pass to year 3*
5 trainees (military doctors): 6 months dispensation for yr 2; 3 months in yr 1; 3 months in yr 3

Table 2 Training process characteristics and poor performance

Training process characteristics	**Total** ***n* = 215**	**Yr 1** ***n* = 215**	**Yr 2** ***n* = 207**	**Yr 3** ***n* = 203**
Employment rate % < 85%		3.3	3.9	5.4
85- 95%		14.4	23.2	36.5
>95%		82.3	72.9	58.1
Illness %: No illness	84.2	95.8	94.2	93.1
1 time	14.4	4.2	5.3	6.4
2 times	1.4	0	0.5	0.5
Mean period of illness in weeks (SD)	10.6 (8.4)	12.1 (9.8)	9.0 (8.4)	10.1 (7.9)
Median	7.5	10.0	6.5	6.0
	n = 34	*n* = 9	*n* = 12	*n* = 14
Maternal leave, *n* women	*n* = 148	*n* = 148	*n* = 142	*n* = 140
1 time	47	18	23	40
2 times	19	0	1	1
Outcome poor performance %	22.7	9.7	13.0	7.4
	n = 49	*n* = 21	*n* = 27	*n* = 15

Table 3 Distribution of shortcomings per competency roles of poor performers per year

	Medical expert			Communicator			Professional			Remaining roles			Years
Trainee#	**Yr 1**	**Yr 2**	**Yr 3**	**Yr 1**	**Yr 2**	**Yr 3**	**Yr 1**	**Yr 2**	**Yr 3**	**Yr 1**	**Yr 2**	**Yr 3**	
001											X		1
003									X				1
006		X			X								1
010	X			X									1
016				X			X						1
021							X						1
023		X	X		X	X		X	X				2
029								X					1
048	X												1
051			X						X				1
056	X	X		X	X		X	X	X				3
059									X				1
062	X												1
063	X			X			X	X	X				3
071		X						X					1
073					X			X					1
076	X	X			X			X		X			2
085	X	X		X			X						2
092								X					1
101		X											1
102		X			X								1
104		X											1
105	X			X			X						1
106					X	X		X	X				2
111									X				1
114									X				1
121	X			X									1
130							X						1
139				X				X	X				3
142				X									1
143								X					1
146				X									1
148								X					1
167								X					1
171								X					1
178					X			X			X		1
185	X												1

	Medical expert			Communicator			Professional			Remaining roles			Years
Trainee#	Yr 1	Yr 2	Yr 3	Yr 1	Yr 2	Yr 3	Yr 1	Yr 2	Yr 3	Yr 1	Yr 2	Yr 3	
189				X			X	X	X				3
190									X				1
195		X											1
200									X				1
205									X				1
210	X			X			X						1
214				X				X					2
215	X	X		X			X						2
216								X					1
222							X			X			1
226									X				1
227								X					1

Table 4 Association between individual characteristics, early assessments and poor performance (*n* = 49) univariate and multivariate logistic regression models in total cohort

		Univariate OR's (95% CI)	**Multivariate OR's (95% CI)**
Individual characteristics	Gender (0:male; 1:female)	0.81 (0.41 – 1.60)	1.29 (0.56 – 2.97)
	Age (in years)	1.18 (1.09 – 1.28)	1.16 (1.06 – 1.27)
	Region medical school (0:NW Europe;1:elsewhere)	6.00 (2.02 – 17.85)	
	Past performance (0: <1 year;1: ≥1year)	1.66 (0.86 – 3.18)	
	Times of application (0:1st time; 1:2nd or 3rd time)	2.46 (1.21 – 5.03)	
	Dispensation	0.84 (0.62 – 1.13)	
	Mean application score	0.09 (0.02 – 0.33)	
Early assessments	Medical expert	0.56 (0.40 – 0.78)	
	Communicator	0.55 (0.38 – 0.80)	0.50 (0.33 – 0.77)
	Professional	0.71 (0.52 – 0.97)	
	Knowledge (0: insufficient; 1: sufficient)	0.15 (0.07 – 0.34)	0.16 (0.07 – 0.40)

OR= Odds ratio; CI= Confidence interval

Table 5 Association between individual. early assessments and training process characteristics and poor performance in respectively first, second and third year (univariate logistic regression)

		First year *n*=215 Univariate OR's (95% CI)	**Second year *n*=207 Univariate OR's (95% CI)**	**Third year *n*=203 Univariate OR's (95% CI)**
Individual characteristic	Gender (0: male; 1: female)	0.71 (0.28 – 1.81)	0.52 (0.23 – 1.19)	1.89 (0.51 – 6.94)
	Age (in years)	1.17 (1.07 – 1.27)	1.15 (1.06 – 1.25)	1.10 (1.00 – 1.21)
	Region medical school (0: NW Europe; 1:elsewhere)	8.22 (2.58 – 26.19)	6.14 (1.94 – 19.40)	0.96 (0.12 – 7.85)
	Past performance (0: <1 year;1: ≥1year)	2.00 (0.77 – 5.17)	2.04 (0.87 – 4.80)	2.67 (0.82 – 8.67)
	Times of application (0: 1st time; 1: 2nd or 3rd time)	1.84 (0.69 – 4.92)	2.30 (0.94 – 5.58)	1.99 (0.63 – 6.31)
	Dispensation	1.41 (0.89 – 2.22)	0.76 (0.52 – 1.11)	0.68 (0.41 – 1.12)
	Mean application score	0.06 (0.10 – 0.39)	0.06 (0.01 – 0.32)	0.25 (0.03 – 1.90)
Early assessments	Medical expert	0.49 (0.32 – 0.77)	0.48 (0.31 – 0.75)	1.03 (0.58 – 1.82)
	Communicator	0.66 (0.41 – 1.07)	0.44 (0.28 – 0.68)	0.94 (0.52 – 1.70)
	Professional	0.58 (0.35 – 0.94)	0.65 (0.42 – 1.01)	1.04 (0.61 – 1.78)
	Knowledge (0: insufficient; 1: sufficient)	0.09 (0.04 – 0.25)	0.28 (0.11 – 0.70)	0.46 (0.14 – 1.56)
Training process characteristics	Employment rate yr 1 (0: <85%; 1: 85-95%; 2:> 95%)	0.73 (0.32 – 1.66)	0.89 (0.39 – 2.03)	0.57 (0.23 – 1.39)
	Illness year 1 (in weeks)		1.23 (1.00 – 1.52)	Not estimable
	Maternity leave year 1 (0: no; 1: yes)		0.82 (0.18 – 3.78)	1.76 (0.36 – 8.57)
	Poor performance year 1		7.16 (2.52 – 20.30)	4.87 (1.36 – 17.44)
	Employment rate yr 2 (0:< 85%; 1: 85-95%; 2: >95%)		0.70 (0.35 – 1.39)	0.40 (0.18 – 0.90)
	Illness year 2 (in weeks)			1.11 (0.94 – 1.31)
	Maternity leave year 2 (0: no; 1: yes)			3.22 (0.93 – 11.10)
	Poor performance year 2			5.57 (1.79 – 17.28)
	Employment rate yr 3 (0: <85%; 1: 85-95%; 2: >95%)			0.50 (0.23 – 1.11)

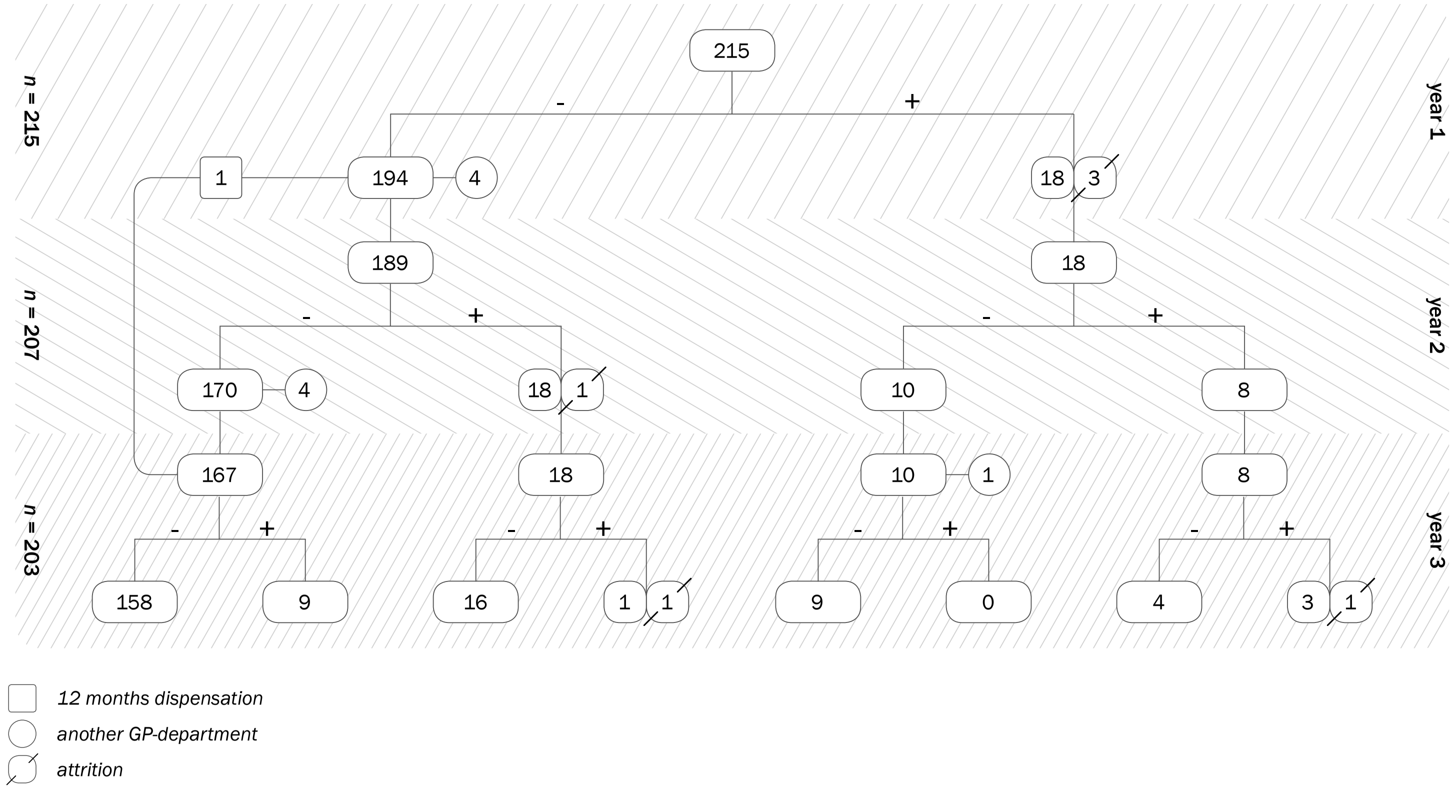
OR= Odds ratio; CI= Confidence interval

Table 5 Association between individual, early assessments and training process characteristics and poor performance in respectively first, second and third year (multivariate logistic regression)

		First year n = 215 Multivariate OR's (95% CI)	**Second year n = 207 Multivariate OR's (95% CI)**	**Third year n = 203 Multivariate OR's (95% CI)**
Individual characteristic	Gender (0: male; 1: female)	1.27 (0.39 – 4.18)	0.82 (0.31 – 2.21)	3.89 (0.83 – 18.11)
	Age (in years)	1.14 (1.02 – 1.26)	1.12 (1.02 – 1.23)	1.10 (0.98 – 1.23)
	Region medical school (0: NW Europe; 1:elsewhere)			
	Past performance (0: <1 year;1: ≥1year)			
	Times of application (0: 1st time; 1: 2nd or 3rd time)			
	Dispensation			
	Mean application score			
Early assessments	Medical expert	0.49 (0.29 – 0.84)		
	Communicator		0.47 (0.30 – 0.75)	
	Professional			
	Knowledge (0: insufficient; 1: sufficient)	0.12 (0.04 – 0.35)		
Training process characteristics	Employment rate yr 1 (0: <85%; 1: 85-95%; 2:> 95%)			
	Illness year 1 (in weeks)			
	Maternity leave year 1 (0: no; 1: yes)			
	Poor performance year 1		4.20 (1.31 – 13.47)	
	Employment rate yr 2 (0:< 85%; 1: 85-95%; 2: >95%)			
	Illness year 2 (in weeks)			
	Maternity leave year 2 (0: no; 1: yes)			
	Poor performance year 2			5.40 (1.58 – 18.47)
	Employment rate yr 3 (0: <85%; 1: 85-95%; 2: >95%)			

OR= Odds ratio; CI= Confidence interval

Figure 1 Flowchart of the trainees during their postgraduate GP Training (*n* = 215); poor performance (+)



Chapter 6

Development of a multi-method selection procedure for postgraduate training based on the CanMEDS in the context of GP training

Tromp F, Vermeulen MI, Mokkink H, Vernooij-Dassen M, Bottema B, Kramer A

submitted

Abstract

Introduction

The selection procedures are frequently more intuitive than theory or evidence-based procedures. Our aim was to present the development of a fair and standardized selection procedure for postgraduate training based on empirical evidence.

Methods

We advocated an assessment procedure with multiple sources of information from various methods to construct an overall judgement by triangulating information across these sources. First, the content of the procedure was determined with a modified Delphi procedure. Then, we selected instruments to be used in the procedure by searching the literature for relevant, feasible, reliable and valid methods.

Results

Consensus on the following CanMEDS roles was reached: ‘medical expert’, ‘communicator’, ‘collaborator’, ‘manager’, and ‘professional’.

Four instruments were included: a knowledge test; a situational judgement test; patterned behaviour descriptive interview, and a series of three work-related simulations.

Conclusion

A competency-based multi-method selection procedure for postgraduate training based on empirical evidence and the CanMEDS framework is available for our training. This procedure will allow for comprehensive standardization that is fair to the candidates. The results can be used during training as a baseline assessment by trainers and candidates. They can employ the scores of the selection instruments to identify future development. Further research is needed to establish the reliability and predictive validity of the procedure. Other medical specialties that utilize the CanMEDS or comparable competency frameworks as a basis for their curriculum could employ this stepwise development model.

Introduction

Selection procedures play a crucial role in obtaining access to medical postgraduate training. These procedures should be credible, fair, and publicly defensible. For many occupational groups, a large body of international research exists investigating best practice selection.[1] In medicine, there is a significant volume of research exploring medical school admission procedures and the link to subsequent performance during medical school. There is relatively little research on developing selection methodology for entry as a trainee to postgraduate training.[2,3] As a result, most specialties continue to select trainees on a subjective and often poorly defined basis.[4] To improve the selection process, Prideaux et al. postulate that selection should be conceptualized as 'assessment for selection'.[3] In doing so, the well-developed quality assurance mechanisms associated with high-stakes assessment can be applied in the selection process.

The first quality assurance mechanism is 'proceeding from a clear blueprint of the content for selection'.[3] The importance of a thorough blueprint has been underlined previously.[5,6] At present, competency frameworks, such as those developed by the Accreditation Council for Graduate Medical Education and the American Board of Medical Specialties (ACGME/ABMS) and the Canadian Medical Education Directives for Specialists (CanMEDS) 2000, guide the construction of curriculums in many countries.[7,8] These competency frameworks constitute the essential abilities that physicians need for optimal functioning.[8] The implementation of these competency frameworks may have consequences for selection, as the competencies can be used for the content of an assessment procedure. They provide all of the qualities for which a specialist should strive as a medical expert, communicator, collaborator, scholar, manager, health advocate, and professional. Not all competencies are appropriate for selection. Some competencies, such as showing empathy, are difficult to teach, and should have been developed already during medical school. Other competencies, such as the execution of specific surgical procedures, can be developed more easily during postgraduate training. This distinction is crucial for determining the selection criteria for each specialty.

The second issue emphasized by Prideaux et al. is that selection should be aligned with the curriculum and assessment. The majority of selection procedures for postgraduate entry are based on cognitive variables, and these variables alone do not adequately predict the performance of competencies such as Professionalism, Communication, or Management.[3,5,6,9] Assessing these competencies during selection may actually be more predictive for success in a postgraduate training than traditional cognitive selection factors, as they are regularly assessed during training. By taking the competency framework as a starting point and assessing the same competencies as in training, the selection procedure is more aligned with curriculum and assessment.

In conceptualizing selection as 'assessment for selection', we can benefit from the findings of experts in the field of assessment and apply these findings to the assessment for selection.[3] Van der Vleuten and Schuwirth state that one instrument is not sufficient in high-stakes assessment.[10] This insight inspired them to advocate programmes of assessment. In a program of assessment multiple sources of information from various methods are used to construct an overall judgement by triangulating information across these sources. Likewise, the selection procedure should consist of several assessment tools and should be considered as the first assessment in a program of assessments. The results of the selection procedure can be seen as a baseline assessment. This procedure enables future trainees to receive feedback at the very begin of the training. Assessments should generate feedback, because feedback promotes learning: it advises trainees regarding observed learning needs; and it motivates trainees to engage in appropriate learning activities.[11]

There have been important developments in the domain of selection for postgraduate training. In the UK, Patterson et al. developed a new selection procedure based on competencies to recruit general practice (GP) trainees.[12] This procedure exhibited predictive validity for future performance during the training of candidates. In the Netherlands, Vermeulen et al. studied the current selection procedure of GP training.[13] This selection procedure is endorsed nationally but conducted locally. It was found that despite the legislation, different standards were used in different institutes and the department itself was a predictor of being admitted. Viewing the results of their study, the authors expressed their doubts about the fairness of the selection procedure and suggested that the current method be reconsidered.

Our aim was to develop a fair, standardized selection procedure for GP training based on empirical evidence and on the leading theoretical studies. Although this procedure was developed in the specific context of GP training, other specialties can benefit from our experiences by describing the process of the development.

Methods

Context of the study

We conducted our study in Dutch Postgraduate GP Training. This training has a nationally endorsed curriculum. The GP departments of the eight university medical centres are responsible for the organization of the three-year postgraduate training. The curriculum is based on the CanMEDS competencies, which are adapted to the specific needs of the specialty.[8] These competencies are assessed during training with the Competency Assessment List (‘Compass’), an instrument that lists the seven competencies.[14] The Compass aggregates the assessments of performance in practice at several points during training. In the development period the number of candidates was decreasing, while the number of vacancies increased. In order to maintain quality of the training, we decided to ‘select out’, meaning that we wanted to identify unsuitable candidates.

Dropouts and poor performers, even if their number might be small, cost the departments substantial effort and money. Their places cannot be filled, which is a waste of resources.

In conclusion, we aimed to develop a selection procedure for postgraduate GP training that

- is able to identify unsuitable candidates;
- is preceded by a content analysis and based on relevant and actual competencies for the specialty;
- exhibits congruity between selection, curriculum, and assessment;
- uses multiple assessment instruments on different levels of Miller with satisfactory predictive validity and reliability; and
- is feasible: the whole procedure has to be executed in not more than one day

Design of the study

Our study consisted of two steps:

Step 1: Establishing the content of the selection procedure

To determine which of the CanMEDS competencies should be targeted for the selection procedure for GP training, we invited a panel of 16 experts, all involved in selecting candidates for GP training. In a two-round process, we asked the panellists to judge 'which of the CanMEDS-competencies should already be present before entering GP training to finish the training successfully'. This inquiry indicates that we aim to identify unsuitable candidates with the new procedure. In cases where one or more of the CanMEDS competencies are missing, candidates will not be admitted to the training, i.e., they will be 'selected out'.

We used the Compass format to determine the content because it lists all seven competencies and their 19 subcompetencies (Box 1). The panel members individually rated the subcompetencies on a nine-point scale, ranging from should not be present before entering GP training (= 1) to should certainly be present before entering GP training (= 9). In the first round, the ratings were made individually at home, with no interaction among the panellists. In the second round, the panel members met under the leadership of a moderator. During the meeting, the panellists discussed their previous ratings, focusing on areas of disagreement. Disagreement among the raters was defined if at least one third of the panel members rated the subcompetency in the range of one to three while at least one third of the other panel members rated the subcompetency treatment in the range seven to nine. After discussing each of the seven competencies, they rerated each subcompetency individually. The two-round process was focused on detecting consensus among the panel members. No attempt was made to force the panel to consensus. The subcompetencies were accepted when they received a mean score of seven or higher.

To corroborate the results of the expert panel and obtain additional information, we organized a focus group meeting with the heads of the training departments. These individuals have a good impression of the problems of non-functioning trainees and dropouts, as they are responsible for the pass-or-fail decisions. The heads of the departments were asked which of the competencies most frequently cause the most significant problems. The panels were not informed about the results of the other panel, and we combined the results of the two panels.

Step 2: Determination of assessment tools

The second step was to determine which instruments should be used to assess the competencies that we found during the first step. In a high-stakes situation, no single assessment instrument can provide the necessary information for judgement.[10,15] We chose to include various instruments that provide different levels of information. Corresponding to the classification that George Miller proposed for the methods of assessment in medical education, we aimed to assess the factual knowledge and performance of the candidates.[15] We assessed whether the candidates were able to act appropriately in a practical situation by exhibiting functional behaviour. This approach complements our endeavour to achieve congruity between the selection, curriculum, and assessment because during training trainees are assessed in a similar manner.

The included instruments should have a good predictive validity and reliability, as confirmed by the literature. We conducted a search using the PubMed database. The following search terms were applied: “internship and residency”, “education, graduate”, “vocational education”, “school admission criteria”. We used the database Psychinfo applying the keywords “personnel selection” with the limitation “ meta-analysis”.

Results

Step 1: Establishing the content of the selection procedure

Of the 16 individuals that we approached, 11 were willing to participate. The panel reached consensus on eight of the 19 subcompetencies in the first, written rating round. In the second round, two subcompetencies of Medical Expertise, one of Communication, three of Collaboration, two subcompetencies of Management, one of Social Accountability, and two of Professionalism were discussed. After hearing positive and negative arguments, the panel re-rated the subcompetencies at the end of the meeting, resulting in consensus on five competencies and nine of their respective subcompetencies. The heads of the departments reported that the competencies Medical Expertise, Communication, Collaboration, and Professionalism caused the greatest difficulty during training. They felt that a significant lack of medical knowledge was an important cause of problems or dropping out during training. The results of the panel of experts and the heads of the departments overlapped to a large extent, and the only competency that was not reported by the department heads was Management, which was contrary to the experts. In box 1, the targeted competencies are printed in italics.

Step 2: Determination of assessment tools
For practical reasons, four instruments were included because we felt that additional instruments made the procedure overly lengthy. The format for each instrument, which describes the competencies that are assessed and evidence from the literature, will be discussed.

The Knowledge Test for General Practice
Because the department heads felt that a significant lack of medical knowledge was an important cause of problems or even dropping out during training, we decided, also for pragmatic and cost efficiency reasons, to assess medical knowledge with the validated National GP Knowledge Test (LHK, Dutch abbreviation) that is used during training to assess the progress of knowledge.[16,17] The knowledge test is based on a blueprint covering all seven competency roles and consists of 120 questions with “correct”/“incorrect”/“do not know” answers.[17] To discourage guessing, the overall score is calculated as the sum of the correct minus incorrect answers and is expressed as a percentage of the maximum score.
Evidence from the literature. Schmidt and Hunter have reviewed the literature on the predictive validity of the various selection instruments and found high predictive validity for knowledge tests.[1] In the medical domain, it was shown that medical knowledge is the basis of performance in practice.[16,18-23] In general, reliability of the LHK varies between 0.60 and 0.76.[17,19-24]

Situational Judgement Test
To assess the ability to use knowledge in a particular context, a situational judgement test (SJT) was included. In a SJT, the candidates are presented with written depictions of professional dilemmas that they may encounter in practice and are asked to identify an appropriate response from a list of alternatives. This test assesses the Professionalism, Management, Collaboration, and Communication competencies.
With the Critical Incidence Technique (CIT) experienced GP’s formulated 20 professional dilemmas for the SJT, each with four alternatives.15 The alternatives can be rated as “very appropriate”, “appropriate”, “neutral”, “inappropriate”, and “extremely inappropriate”. An example is provided in Box 2.
Evidence from the literature. The studies of Lievens and Patterson have demonstrated good predictive validity of the SJT for future performance.[24-26] In a meta-analysis it was

found that the reliability of SJTs ranged from 0.43 to 0.94.[27] In selection for postgraduate training in general practice in the UK, internal consistency of the SJT ranged from 0.80 to 0.83.[28]

Work-related simulations

To assess the candidates' ability to act appropriately in a practical situation, work-related simulations (SIM) were included. The SIM provide the candidates a good impression of daily practice. The Medical Expertise, Communication, Management, Collaboration and Professionalism competencies are assessed with this exercise.

The scripts were developed, also using the CIT, and an example is provided in Box 3. These SIM apply the same principles as the Multiple Mini Interview (MMI) and the Objective Structured Clinical Examination (OSCE).[29] Both MMI and OSCE provide a series of short testing stations and have shown to have superior reliability to a single long case.[29,30] To reduce the context specificity, we developed three short simulations. The observers participated in one day of training in behavioural observation and rating.

Evidence from the literature. Work-related simulations show satisfactory levels of reliability, content- and predictive validity.[12,28,29,31]

Patterned Behaviour Descriptive Interview

The fourth assessment instrument that we chose was the Patterned Behaviour Description Interview (PBDI), which is widely used in selection.[32] The PBDI is based on the premise that past behaviour predicts future behaviour.[32] The interview focuses on the evaluation of reactions in actual situations from each candidates' past, relevant to the targeted competencies. The assessed competencies are 'Communication', 'Management', 'Collaboration', and 'Professionalism'. The interviewers were trained in the technique of this specific format of interviewing and how to rate the candidates' answers.

Evidence from the literature. Schmidt and Hunter have demonstrated that good validity coefficients are found for the more structured and systematic techniques, such as patterned behaviour descriptive interviews.[1] In the medical literature, we found that the use of behaviour-specific questions during the interview improved the predictive validity and reliability.[19,33,34]

In summation, our procedure consisted of

- The Knowledge Test for General Practice
- A Situational Judgement Test
- Work-related simulations
- A Patterned Behaviour Descriptive Interview

Table 1 summarizes which competencies are assessed by each of the four instruments.

Discussion

A competency-based selection procedure for postgraduate training based on empirical evidence and theoretical insight is available for postgraduate training in GP. We developed this procedure guided by empirical data and the recommendations of the leading experts in the field.[2,3,6,9,12,35]

This procedure allows for comprehensive standardization that should be more fair to the candidates than the existing methods.[2,13] To ensure good standardization in the future, all of the GP departments should work together to produce procedural and best practice manuals and ensure that all assessors and simulators have been properly trained for the best practice selection procedures, as is the case in the UK. The involvement of all departments is crucial and was the key factor for success in the UK in developing a standardized national selection procedure.[2]

To our knowledge, the developed selection procedure is the first to be based on the CanMEDS. One of the characteristics of the procedure is a job analysis to classify the core and specific competencies, as recommended by Prideaux.[3] With the aid of experts in the field, we decided which competencies should be targeted for selection. These generic formulated competencies were translated to behavioural indicators using the CIT.[15] By assessing the candidates according to the content of the curriculum, the candidates are confronted with what will be expected of them during training.[3,5,6] With this transparency, the candidates are likely to develop a realistic perception of the job role. This assessment may potentially reduce the number of false positives, thereby reducing attrition rates and problems occurring during training because candidates would have a more realistic insight of what the job entailed before enlisting.[6]

To determine the predictive validity of various instruments, we found that the results were not univocal. There seems to be a considerable debate about the predictive validity of each single instrument. In our procedure however, we do not use one single instrument. No single instrument can assess all competencies, Moreover, one single assessment has limitations, such as case-specificity, or low reliability.[36] The four instruments we chose provide information of the candidates on different levels. Our procedure will incorporate several competency elements and multiple sources of information to evaluate those competencies on multiple occasions using credible standards. The information obtained will have to be aggregated into a final decision.

Trainers and trainees can use the scores of the selection instruments to identify future development needs for candidates. The procedure can be considered a baseline assessment. It enables future trainees to receive feedback before the training begins. With the aid of this feedback, the future trainees are able to remedy potential shortcomings in the earliest stage of the training. The ability to provide and receive feedback at such an early stage of training is unique. The rejected candidates will receive feedback enabling them to work on their deficiencies, thus giving them a fair chance if they choose to apply again.

Although our study was based on developments in the UK our procedure has one important difference.[6,9,12] We deliberately included the PBDI because it is widely used in selection and has good predictive validity.[1]

Having established a theory-driven and evidence-based selection procedure, the next step will be to evaluate its effects and costs. The reliability and predictive validity of the proposed procedure will be assessed by conducting a longitudinal study, and the candidate's perspective and perceptions of fairness must be considered.[37] Further work exploring how to utilize the output from the selection process to inform personal development planning is important.

Strength and limitations

High numbers of candidates for places in medical training could prove to be a substantial investment of time and money and so jeopardize feasibility. One solution to unburden the procedure could be to use the Knowledge Test and the SJT as preselection tools, as in the UK.[2,31]

We did not include former academic performance as a selection variable. Although there is some criticism among researchers that the educational curricula and quality of teaching may differ among institutions,[38] academic performance can be considered as a good predictor of future performance; this consideration will be investigated in the future.[5,38-40]

Conclusion

We developed a selection procedure based on the CanMEDS using relevant, feasible, reliable and valid instruments assessing multiple competencies at three levels of Miller's pyramid. The results of the selection procedure can be used at the start of the training as a baseline assessment. The next step will be to study the instrument's reliability, predictive validity fairness, costs, and future use in training, as well as the candidates' reactions.

This study described a stepwise process for the development of a competency-based selection procedure. Other medical specialties that utilize the CanMEDS or a comparable competency framework as a basis for their curriculum could employ this development model.

References

1. Schmidt FL, Hunter JE. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin 1998;124:262-74.
2. Plint S, Patterson F. Identifying critical success factors for designing selection processes into post-graduate specialty training: the case of UK general practice. Postgrad Med J 2010;86:323-7.
3. Prideaux D, Roberts C, Eva K, Centeno A, McCrorie P, McManus C, et al. Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 Conference Med Teach 2011;33:215-23.
4. Thordarson DB, Ebramzadeh E, Sangiorgio SN, Schnall SB, Patzakis MJ. Resident selection: how we are doing and why? Clin Orthop Relat Res 2007;459:255-9.
5. Lee AG, Golnik KC, Oetting TA, Beaver HA, Boldt HC, Olson R, et al. Re-engineering the resident applicant selection process in ophthalmology: a literature review and recommendations for improvement. Surv Ophthalmol 2008;53:164-76.
6. Patterson F, Ferguson E, Thomas S. Using job analysis to identify core and specific competencies: implications for selection and recruitment. Med Educ 2008;42:1195-204.
7. Horowitz SD, Miller SH, Miles PV. Board certification and physician quality. Med Educ 2004;38:10-1.
8. Frank JR, Jabbour M, Tugwell p. Skills for the new millenium: Report of the societal needs working group, CanMEDS 2000 Project. Annals of the Royal College of Physicians and Surgeons of Canada 1996;29:206-16.
9. Patterson F, Ferguson E, Lane P, Farrell K, Martlew J, Wells A. A competency model for general practice: implications for selection, training, and development. Br J Gen Pract 2000;50:188-93.
10. van der Vleuten CPM, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005;39:309-17.
11. Shepard LA. The Role of Assessment in a Learning Culture. Educational Researcher 2000;29:4-14.
12. Patterson F, Ferguson E, Norfolk T, Lane P. A new selection system to recruit general practice registrars: preliminary findings from a validation study. BMJ 2005;330:711-4.
13. Vermeulen MI, Kuyvenhoven MM, Zuithoff NP, Tromp F, van der GY, Pieters RH. Selection for Dutch postgraduate GP training; time for improvement. Eur J Gen Pract 2012;18:201-5.
14. Tromp F, Vernooij-Dassen M, Grol R, Kramer A, Bottema B. Assessment of CanMEDS roles in post-graduate training: The validation of the Compass. Patient Educ Couns 2012;89:199-204.
15. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990 Sep;65(9 Suppl):S63-S67.
16. Kramer AW, Zuithoff P, Jansen JJ, Tan LH, Grol RP, van d, V. Growth of self-perceived clinical competence in postgraduate training for general practice and its relation to potentially influencing factors. Adv Health Sci Educ Theory Pract 2007;12:135-45.
17. van Leeuwen YD, Mol SS, Pollemans MC, Drop MJ, Grol R, van d, V. Change in knowledge of general practitioners during their professional careers. Fam Pract 1995;12:313-7.
18. Boyse TD, Patterson SK, Cohan RH, Korobkin M, Fitzgerald JT, Oh MS, et al. Does medical school performance predict radiology resident performance? Acad Radiol 2002;9:437-45.
19. Brothers TE, Wetherholt S. Importance of the faculty interview during the resident application process. J Surg Educ 2007;64:378-85.
20. Carmichael KD, Westmoreland JB, Thomas JA, Patterson RM. Relation of residency selection factors to subsequent orthopaedic in-training examination performance. South Med J 2005;98:528-32.
21. Dirschl DR, Dahners LE, Adams GL, Crouch JH, Wilson FC. Correlating selection criteria with subsequent performance as residents. Clin Orthop Relat Res 2002;265-71.
22. Ram P, van d, V, Rethans JJ, Schouten B, Hobma S, Grol R. Assessment in general practice: the predictive value of written-knowledge tests and a multiple-station examination for actual medical performance in daily practice. Med Educ 1999;33:197-203.
23. Thundiyil JG, Modica RF, Silvestri S, Papa L. Do United States Medical Licensing Examination (USMLE) scores predict in-training test performance for emergency medicine residents? J Emerg Med 2010;38:65-9.
24. Lievens F, Patterson F. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. J Appl Psychol 2011; 96:927-40.

25. Patterson F, Ashworth V, Zibarras L, Coan P, Kerrin M, O'Neill P. Evaluations of situational judgement tests to assess non-academic attributes in selection. Med Educ 2012;46:850-68.
26. Patterson F, Ashworth V, Mehra S, Falcon H. Could situational judgement tests be used for selection into dental foundation training? Br Dent J 2012;213:23-6.
27. McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. Use of situational judgment tests to predict job performance: a clarification of the literature. J Appl Psychol 2001;86:730-40.
28. Patterson F, Baron H, Carr V, Plint S, Lane P. Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. Med Educ 2009;43:50-7.
29. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. Med Educ 2004; 38:314-26.
30. Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. Predictive validity of the multiple mini-interview for selecting medical trainees. Med Educ 2009;43:767-75.
31. Irish B, Patterson F. Selecting general practice specialty trainees: where next? Br J Gen Pract 2010;60:849-52.
32. Huffcutt AI, Conway JM, Roth PL, Stone NJ. Identification and meta-analytic assessment of psychological constructs measured in employment interviews. J Appl Psychol 2001;86:897-913.
33. Altmaier EM, Smith WL, O'Halloran CM, Franken EA, Jr. The predictive utility of behavior-based interviewing compared with traditional interviewing in the selection of radiology residents. Invest Radiol 1992;27:385-9.
34. Wood PS, Smith WL, Altmaier EM, Tarico VS, Franken EA, Jr. A prospective study of cognitive and noncognitive selection criteria as predictors of resident performance. Invest Radiol 1990;25:855-9.
35. Patterson F, Carr V, Zibarras L, Burr B, Berkin L, Plint S, et al. New machine-marked tests for selection into core medical training: evidence from two validation studies. Clin Med 2009;9:417-20.
36. van der Vleuten CPM, Schuwirth LW, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. Best Pract Res Clin Obstet Gynaecol 2010;24:703-19.
37. Patterson F, Zibarras L, Carr V, Irish B, Gregory S. Evaluating candidate reactions to selection practices using organisational justice theory. Med Educ 2011;45:289-97.
38. Tanilon J, Segers M, Vedder P, Tillema H. Development and validation of an admission test designed to assess samples of performance on academic tasks. Studies In Educational Evaluation 2009;35:168-73.
39. Fine PL, Hayward RA. Do the criteria of resident selection committees predict residents' performances? Acad Med 1995;70:834-8.
40. Siu E, Reiter HI. Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. Adv Health Sci Educ Theory Pract 2009;14:759-75.

Box 1 Overview of the seven competencies

Subcompetencies that are targeted for selection are printed in *italics*.

Medical Expertise
Interprets symptoms in context
Applies the diagnostic, therapeutic, and preventive arsenal of the profession in an appropriate and evidence-based manner
Provides primary care in a systematic manner

Communication
Develops effective treatment relationships with patients
Applies communication techniques and resources appropriately
Ensures that the patient is actively involved in the decision making

Collaboration
Contributes to effective intra- and interdisciplinary collaboration
Applies collaboration skills appropriately
Makes appropriate referrals on the basis of a current insight into the expertise of other care providers

Management
Provides integral and appropriate general practice care that is continuous and accessible
Applies organizational and management principles appropriately
Uses information technology for optimal patient care

Social Accountability
Promotes the health of individual patients and groups of patients
Acts in accordance with the legislation that applies to the general practitioner

Science and Education
Underpins care in an academically sound manner
Promotes the expertise of students, trainees, colleagues, and other care providers

Professionalism
Maintains a balance between personal and professional roles
Works systematically and purposefully to improve his or her professional performance
Addresses differences in standards and values consciously within the context of professional ethics

Table 1 List of the four instruments indicating which subcompetencies they assess.

	LHK	SJT	SIM	PBDI
Medical Expertise				
Interprets symptoms in context	X		X	
Applies the diagnostic, therapeutic, and preventive arsenal of the profession in an appropriate and evidence-based manner	X			
Communication				
Develops effective treatment relationships with patients		X	X	X
Applies communication techniques and resources appropriately		X	X	
Collaboration				
Applies collaboration skills appropriately		X	X	X
Management				
Applies organizational and management principles appropriately		X	X	X
Professionalism				
Maintains a balance between personal and professional roles		X	X	X
Works systematically and purposefully to improve one's professional performance		X		X
Addresses differences in standards and values consciously within the context of professional ethics		X	X	X

Box 2 Example of a situation with four behavioural options in the SJT

A patient encounters his GP with complaints of fatigue. Lately, the man has worked hard under a lot of pressure; he is worried that something physically is wrong. After thorough examination the GP tells him that worry is not needed. His complaints are most probably a consequence of his hard working. Patient: "Still, I don't trust it. Could you arrange more medical examinations in the hospital?"

Reactions of the GP:

1. OK, but if I explain to you that these complaints are normal in your situation, why do you want more examinations? What do you think will be the advantage for you?
2. OK, if you are still so worried and I cannot take away your concerns, then perhaps it is better to refer you to internal medicine.
3. You are still not convinced? OK, let's see why you are so uncertain. Maybe I can reassure you.
4. You are still worried. I don't understand why. I think you are somewhat exaggerating.

Box 3 Example of a script of a work-related simulation

Instruction:

You are a GP working in a group practice. The surgery hours are very busy today. You are already overrun by 30 min. The next consultation with Mrs. R. She has been a patient in your practice for only one year, so you don't know her very well. Mrs. R indicated that her reason for the encounter is headache. This instance is not the first time she consults you for this reason. From her medical file, you know she saw a neurologist 18 months ago. She was diagnosed as having pain that originated from muscle tension. She has had physiotherapy with varying success. Mrs. R is married and has two children.

Remember: you only have 10 minutes for the consultation. The waiting room is full, and you are already 30 min late. Your assistant has urged you to hurry and make up for lost time because one of your colleagues had to leave suddenly for personal reasons. You have to take over some of his patients as well. Your assistant also warned you that Mrs. R can be very long-winded.
Physical examination is not necessary. You may assume that a physical examination will not yield further information.

Chapter 7

A competency based selection procedure for Dutch postgraduate GP training; a pilot study on validity and reliability

Vermeulen MI, Tromp F, Zuithoff NPA, Pieters HM, Damoiseaux RAMJ, Kuyvenhoven MM

Eur J Gen Pract 2014; Early Online: 1-7 doi:10.3109/13814788.2014.885013

Abstract

Background

Historically, semi-structured interviews (SSI) have been the core of the Dutch selection for postgraduate general practice (GP) training. This paper describes a pilot-study on a newly designed competency-based selection procedure that assesses whether candidates have the competencies that are required to complete GP training.

Objectives

The objective was to explore reliability and validity aspects of the instruments developed.

Methods

The new selection procedure comprising the National GP Knowledge Test (LHK), a situational judgement tests (SJT), a patterned behaviour descriptive interview (PBDI) and a simulated encounter (SIM) was piloted alongside the current procedure. Forty-seven candidates volunteered in both procedures. Admission decision was based on the results of the current procedure.

Results

Study participants did hardly differ from the other candidates. The mean scores of the candidates on the LHK and SJT were 21.9% (SD 8.7) and 83.8% (SD 3.1), respectively. The mean self-reported competency scores (PBDI) were higher than the observed competencies (SIM): 3.7 (SD 0.5) and 2.9 (SD 0.6), respectively. Content-related competencies showed low correlations with one another when measured with different instruments, whereas more diverse competencies measured by a single instrument showed strong to moderate correlations. Moreover, a moderate correlation between LHK and SJT was found. The internal consistencies (intraclass correlation, ICC) of LHK and SJT were poor, while the ICC of PBDI and SIM showed acceptable levels of reliability.

Conclusion

Findings on content validity and reliability of these new instruments are promising to realize a competency based procedure. Further development of the instruments and research on predictive validity should be pursued.

Introduction

As in most European countries, semi-structured interviews are the core of the selection procedure for Dutch postgraduate general practice (GP) training.[1] However, this instrument weakly predicts future performance.[2] In addition to this shortcoming, admission decisions depend on the selection processes of the respective eight Dutch training departments, although these should be equally carried out according to national regulations.[1] This indicates that a candidate who is admitted to one department may be rejected by another, which generates doubts regarding fairness of the procedure.

In 2005, a competency-based curriculum was introduced, the development of a competency-based selection procedure, like in the UK and Denmark, had to be developed, and aimed to assess whether candidates have competencies needed to complete training successfully.[3,4] Such a procedure is based on the principle that high-stake decisions should be made after consulting assessments from multiple sources and a variety of instruments.[4,5] Patterson et al., showed auspicious results in reliability and validity of a multiple procedure for GP training in the UK.[3,6-8]

For the Netherlands a comparable procedure has been developed. GP experts selected the essential competencies in a Delphi procedure from four role domains; ‘medical expertise’, ‘communication’, ‘collaboration’ and ‘professionalism’.[9] A critical appraisal of the literature resulted in the selection of four instruments: a knowledge test, a situational judgement test, a patterned behaviour descriptive interview and a simulated encounter.[3,4,6-8,10-18] They intended to assess candidates’ ability to cope with complex tasks and to manage (medical) problems in general practice on different levels of Miller’s pyramid, namely: ‘knows’, ‘knows how’ and ‘shows how’ (Table 1).[19]

To explore the content validity and reliability of the instruments, 47 candidates for GP training volunteered to complete the new competency-based procedure alongside the current procedure to answer the following questions:

- How are assessments of content-related and divergent competencies associated between and within the instruments?
- What is the internal consistency of the instruments?
- How do candidates evaluate the relevance and fairness of the instruments?

Methods

Study design

All candidates in the selection procedure of April 2011 for the postgraduate GP training in Nijmegen and Utrecht, received written information regarding the goal and process of the study and were invited to complete the new competency-based selection procedure. Candidates who were willing to participate signed an informed consent form. These volunteers received written feedback on their performance in the new instruments for their private use and received a gift voucher of €50. The current procedure determined the admission decision. Both procedures were carried out independently from each other. The study was executed according to the code conduct for the use of personal data in scientific research.[20]

Data collection

Individual characteristics. Data of all candidates on gender, age (in years), past performance - clinical experience as medical doctor (< one year; ≥ one year), area of medical school (NW Europe; elsewhere), the number of times applied (the first time; > the first time) and the admission decision into GP training were extracted from administrative databases and clerically anonymized before data processing.[20]

Current procedure

Semi-structured interview (SSI). To assess personal qualities of the candidates, interviews were conducted by a selection committee consisting of a staff member, trainer and trainee after a tailored instruction.[1,21] Each member of the selection committee independently assessed the candidates' motivation, orientation on the job, learning needs and personal attributes. In Nijmegen, a four-point scale was used - insufficient (0); uncertain (1); sufficient (2); good (3); in Utrecht, a three-point scale - below (1); upon (2); above standard (3). The Nijmegen scores of '0' and '1' were recoded to correspond to the Utrecht score of '1.' The final score for each quality was the mean score of the three assessors. The head of the department decided on admission, taking the scores and underpinning of the members of selection committee into account.

New procedure

Knowledge Test for General Practice (LHK, Dutch abbreviation). To assess medical knowledge, the national GP knowledge test was used.[17,22] Twice a year, GP experts develop a completely new LHK, consisting of case-vignettes addressing the entire GP domain, with multiple choice questions. The sum of the correct answers minus the sum of the incorrect answers - corrected for guessing - was converted into a percentage score.[17,22]

Situational Judgement Test (SJT). to measure candidates' cognitions about how to effectively resolve practical dilemmas in a GP-setting, an online SJT was developed by highly experienced GPs to assure face validity.[6,12] This test consists of 20 situations with four behavioural options each (Box 1). The candidates indicated the extent to which they perceived the 80 presented options to be effective (1 = completely ineffective, 5 = completely effective). Subsequently, the extent to which the scores corresponded to those of 15 experienced GPs was determined by calculating the absolute difference between the candidates' scores and the mean score of the experienced GPs.

The absolute difference was subtracted from the maximum score and converted into a percentage score.

Patterned behaviour descriptive interview (PBDI). In the PBDI, candidates were asked to report on clinical experiences in which they showed: 'empathy', 'collaboration', 'coping with pressure', 'respect', 'self-care' and 'reflection'. Two staff members explored candidates' behaviour during those situations using the STAR (S = situation, T = task, A = action, R = result) technique.[13] Each competency was scored by consensus on a five-point scale (1 = absent, 2 = doubtful, 3 = average, 4 = sufficient and 5 = good) with written substantiation. All assessors attended a one-day training session to familiarize them with the STAR technique and the competency assessment.

Simulated encounter (SIM). The SIM aimed to measure three competencies: 'medical expertise', 'doctor-patient communication' and 'professionalism'.[3] Experienced GPs constructed two cases: 'a patient with heart complaints' and 'a patient with dyspnoea'. Candidates worked through one, randomly chosen, case. Assessors and actors attended a one-day training as well. Competencies were assessed by one assessor on a five-point scale (see PBDI) with written substantiation.

Deliberation. The assessments of all instruments were aggregated and validated in a deliberation session to evaluate assessments of candidates under thresholds. Threshold for the LHK and SJT was the mean score of the group of candidates minus one SD. The threshold for the PBDI and SIM assessments was a score of '1' (absent) on at least one competency or a score of '2' (doubtful) on at least two competencies.

Evaluation. The candidates indicated on a precoded response sheet the degree (ranging from 1 'strongly disagree' to 5 'strongly agree') to which they considered the content of the instruments relevant to general practice, the degree to which the instrument allowed them to demonstrate their competence and whether or not they considered the instrument fair for all candidates (yes/no).

Data analysis

First, individual characteristics of the study population and the remaining candidates were explored.[23] Assessments of competencies, qualities and instruments were presented as descriptive statistics (mean; SD). Few data were missing (2,8% of all data); 75% of these missing's belonged to the SSI data. In those cases, mean values of the group of candidates were imputed.[24]

Next, we investigated to which degree content-related competencies or qualities measured by different instruments and divergent competencies or qualities measured by the same instruments were associated. The associations between the (overall) instrument scores were estimated. All associations were expressed as Pearson's correlation coefficient; all rating scales were approximately continuous.[1] The internal consistency of the instruments was estimated using intraclass correlation (ICC).[25]

Finally, we compared the mean scores (SD) of the candidates' evaluation of the new instruments with the SSI evaluation scores.[23]

The analyses were performed using SPSS version 20.

Results

Study-population

In Nijmegen and Utrecht 60 and 63 candidates, respectively, participated in the current procedure. Forty-eight candidates were willing to participate in both procedures; one candidate withdrew due to personal circumstances. The study-population (*n* = 47) did not differ from the other candidates (*n* = 76) regarding to gender, age, number of times applied, past performance and percentage of admission, except region of medical school (Table 2). Ten of the 47 candidates were rejected (21%).

Candidates' assessments

In the SSI, motivation, orientation on the job, learning needs and personal attributes were assessed about as high, above the '2' scale point on a three-point scale (Table 3). Four, five, three, and five candidates, respectively, were assessed as 'below standard' for these qualities.

The mean score of the LHK was 21.9% (SD 8.7). Ten candidates scored below the threshold. The mean difference score of the SJT was 83.8% (SD 3.1); the variance was nearly three times less than the variance of the mean LHK score. Six candidates scored below the threshold.

The mean scores of self-reported competencies assessed by the PBDI were higher than the observed competencies assessed by the SIM on a five-point scale; PBDI 3.7 (SD 0.5); SIM 2.9 (SD 0.6). The assessors used all scale points; thus, it was possible to assess differences between candidates. Six candidates were selected for deliberation based on the PBDI, and 12 on the SIM.

Overall, 26 candidates (55.3%) were eligible for deliberation; 19 candidates scored below the threshold on one instrument, six candidates on two instruments and one candidate on three instruments.

Associations

Gender was weakly associated with motivation, learning needs and personal attributes in the current procedure (r: 0.34, 0.41, 0.32, respectively); women received moderately better scores on these qualities. Age showed a weak negative association with learning needs (r: – 0.30). The remaining individual characteristics were not associated with personal qualities in the current procedure (not in the table).

In general, content-related competencies measured by different instruments were weakly associated with one another, whereas more divergent competencies measured by a single instrument were moderately to strongly associated (Table 4). For example, there was a weak correlation (r: 0.28) between empathy in the PBDI and doctor-patient communication in the SIM. However, there was a strong correlation between professionalism and doctor-patient communication in the SIM (r: 0.76). Therefore, the overall PBDI and SIM scores were considered to be relevant indicators of the self-reported or general clinical competencies.

Individual characteristics were not correlated with knowledge (LHK), with cognitions on effectively coping (SJT), with self-reported (PBDI) or shown competencies (SIM) (Table 5). Gender was correlated with the overall score of the SSI; female candidates scored higher. There was a moderate correlation between the overall score for personal qualities measured by the SSI and general clinical competencies measured with SIM (r: 0.34), but not with the remaining instruments. A moderate correlation (r: 0.34) was found between LHK ('knows' level) and SJT ('knows-how' level). No correlations were found between the remaining instruments.

Reliability

The internal consistency of the SSI based on the four personal quality scores was moderate (ICC: 0.65). The internal consistencies of the LHK and SJT were poor (ICC: 0.59 and 0.55, respectively). The internal consistencies of the PBDI and SIM showed acceptable levels of reliability (ICC: 0.79 and 0.73, respectively).

Candidates' evaluation

Candidates perceived the content of the LHK as the most relevant to GP training. Compared to the SSI, the SJT, PBDI and SIM assessments were perceived as providing the same or better opportunities for demonstrating GP competence and as being more fair to candidates (Table 6).

Discussion

Main findings

Scores of competencies measured with instruments on different levels of Miller's pyramid were hardly associated with each other. Two instruments (PBDI and SIM) showed acceptable reliability; the ICC of the LHK and SJT was poor. In general, the candidates considered the instruments used in the new procedure to be relevant and more fair than the SSI in the current procedure; a similar result was obtained in the UK.[8]

Strengths and limitations

Forty-seven candidates completed both procedures, allowing us to compare the results of the assessments. Face validity was guaranteed, because the selection of the competencies, the development of the four instruments and the assessments were performed by experienced GPs. However, the study has some limitations. First, the participants were volunteers; therefore, they are likely to have endorsed the new procedure more positively and participants did not represent candidates who had finished medical school outside of north western Europe. In addition, the correlation between qualities and competencies found in the SSI, PBDI and SIM appeared to contribute to a relatively high ICC but could be attributable to a halo effect. Investigating this possibility was beyond the scope of this study.

Interpretation

The new procedure was based on competencies, selected by experts in a Delphi procedure, and theoretical considerations regarding assessment programs.[3-5,26] Therefore, we assume this new procedure to be more relevant and suitable. The LHK and SJT are easily to be implemented. The PBDI and SIM require well-trained assessors and actors.

It is important to assess medical knowledge before the training, while knowledge is the basis of Miller's pyramid, and a low level of knowledge is a predictor of poor performance.[2,11,18] The choice of using the LHK, a validated and rather reliable test for trainees, was rather pragmatic, as in the Netherlands different progress tests during medical school are administered.[17,27] In general reliability of the LHK varies between 0.60 and 0.76.[27] The reliability of the LHK used in this study was lower, which may be a result of the fact that we did not perform an item analysis resulting in deleting unsuitable items or the fact that the population was different from the trainees. The Grade Point Average of the master, including cognitive and non-cognitive assessments, may become more relevant in the near future.[2,4]

Compared to the SJT used in the selection of GP training in the UK, we found disappointing psychometric results: small variance and poor reliability.[6,12] Development with more situations may improve the instrument. Further testing is needed to determine the value of the SJT in the Dutch context.

The PBDI and SIM have relatively high ICCs, suggesting that both assessments are suitable for high-stakes decisions. Given the perspective that reporting on competencies is easier than ‘showing how,’ the generally higher scores of the candidates on the PBDI were expected.

The assessments of GP knowledge, cognitions on coping with dilemmas, self-reported and observed competencies were found to be weakly correlated with one another, which could be attributed to the different levels that were assessed. Approximately half of the candidates were eligible for deliberation, suggesting that all ‘borderline’ candidates could be discussed to reach a fair and balanced admission decision.

Implications for the future

All of the instruments tested can be implemented and should be evaluated in the context of the entire procedure. The costs of using the LHK and SJT are relatively low. However, the SJT should undergo further development to improve reliability. The PBDI and SIM require extensive resources and logistics to implement. The known weaknesses of rater-based assessments, including the halo effect, may be reduced by expanding the number of assessors. Integrating more mini-interviews into the procedure could prevent a bias caused by context specificity.[5,16,26,28] The narrative feedback provided on the LHK, PBDI and SIM assessments was unexpectedly rich and could be used at the beginning of GP training to establish an individual development plan for each trainee.

Conclusion

The four complementary instruments of the new procedure were found promising and applicable in a competency based procedure. Further development of the instruments and research on their predictive validity should be pursued to establish a relevant and publicly defensible selection for GP training.

References

1. Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, Tromp F, Graaf van der Y, Pieters HM. Selection for Dutch postgraduate GP training; time for improvement. Eur J Gen Pract 2012;18:201-5.
2. Siu E, Reiter HI. Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. Adv in Health Sci Educ 2009;14:758-75.
3. Patterson F, Ferguson E, Norfolk T, Lane P. A new selection system to recruit general practice registrars: preliminary findings from a validation study. BMJ 2005;330:711-4.
4. Prideaux D, Roberts C, Eva K , Centeno A, McCrorie P, McManus C et al. Assessment for selection for the health care professions and specialty training: Consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach 2011;33:215-23.
5. van der Vleuten CPM, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005;39:309-17.
6. Patterson F, Ashworth V, Zibarras L, Coan P, Kerrin M, O'Neill P. Evaluations of situational judgement tests to assess non-academic attributes in selection. Med Educ 2012;46:850-68.
7. Irish B, Patterson F. Selecting general practice specialty trainees: where next? Br J Gen Pract 2010;60:849-52.
8. Patterson F, Zibarras L, Carr V, Irish B, Gregory S. Evaluating candidate reactions to selection practices using organisational justice theory. Med Educ 2011;45:289-97.
9. Tromp F, Vernooij-Dassen M, Grol R, Kramer A, Bottema B. Assessment of CanMEDS roles in postgraduate training: The validation of the Compass. Patient Educ Couns 2012;89:199-204.
10. Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. Predictive validity of the multiple mini-interview for selecting medical trainees. Med Educ 2009;43:767-75.
11. Schmidt FL, Hunter JE. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin 1998;124:262-74.
12. Lievens F, Patterson F. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. J Appl Psychol 2011;96:927-40.
13. Huffcutt AI, Conway JM, Roth PL, Stone NJ. Identification and meta-analytic assessment of psychological constructs measured in employment interviews. J Appl Psychol 2001;86:897-913.
14. Altmaier EM, Smith WL, O'Halloran CM, Franken EA, Jr. The predictive utility of behavior-based interviewing compared with traditional interviewing in the selection of radiology residents. Invest Radiol 1992;27:385-9.
15. Brothers TE, Wetherholt S. Importance of the faculty interview during the resident application process. J Surg Educ 2007;64:378-85.
16. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. Med Educ 2004;38:314-26.
17. Ram P, van der Vleuten C, Rethans JJ, Schouten B, Hobma S, Grol R. Assessment in general practice: the predictive value of written-knowledge tests and a multiple-station examination for actual medical performance in daily practice. Med Educ 1999;33:197-203.
18. Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, van der Graaf Y, Pieters HM. Attrition and poor performance in general practice training; age, competence and knowledge play a role. Ned Tijdschr Geneeskd 2011;155:A2780.
19. Miller GE. The assessment of clinical skills/ competence/ performance. Acad Med 1990;65:S63-7.
20. Association of Universities the Netherlands. (In Dutch: Gedragscode voor gebruik van persoonsgegevens in wetenschappelijk onderzoek) 2005.
21. Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, Graaf van der Y, Damoiseaux RAMJ. Dutch postgraduate GP selection procedure; reliability of interview assessments. BMC Fam Pract 2013, 14:43 doi:10.1186/1471-2296-14-43.
22. van Leeuwen YD, Pollemans MC, Mol SSL, Eekhof JAH, Grol R, Drop MJ. Dutch knowledge test for general practice: issues of validity. Eur J Gen Pract 1995;1:113-7.
23. Gardner MJ, Altman DG. Statistics with confidence. Confidence intervals and statistical guidelines. London: BMJ 1989.

24. van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J of Clin Epidemiol 2006;59:1102-9.
25. McGraw KO, Wong SP. Forming inferences about some intraclass correlations coefficients. Psychological methods 1996;1:30-46.
26. Schuwirth LWT, Vleuten van der CPM. How to design a useful test: the principles of assessment. In: Swanwick T. edited. Understanding Medical Education: Evidence, Theory and Practice. First ed. London: Wiley-Blackwell, 2010;195-207.
27. Kramer AWM, Düsman H, Tan LHC, Jansen KJM, Grol RPTM and van der Vleuten CPM. Effect of extension of postgraduate training in general practice on the acquisition of knowledge of trainees. Fam Pract 2003;20:207-12.
28. Gingerich A, Regehr G, Eva KW, Rater-based assessments as social judgements: rethinking the etiology of raters errors. Acad Med 2011;86:S1-7.

Box 1 Example of a situation with four behavioural options in the SJT

A patient encounters his GP with complaints of fatigue. Lately, the man has worked hard under a lot of pressure; he is worried that something physically is wrong. After thorough examination the GP tells him that worry is not needed. His complaints are most probably a consequence of his hard working. Patient: "Still, I don't trust it. Could you arrange more medical examinations in the hospital?"

Reactions of the GP:

1. OK, but if I explain to you that these complaints are normal in your situation, why do you want more examinations? What do you think will be the advantage for you?
2. OK, if you are still so worried and I cannot take away your concerns, then perhaps it is better to refer you to internal medicine.
3. You are still not convinced? OK, let's see why you are so uncertain. Maybe I can reassure you.
4. You are still worried. I don't understand why. I think you are somewhat exaggerating.

Table 1 Domains, Millers' pyramid levels and instruments of the current and the new competency based selection procedure

Domains	Millers' level	Instrument; personal qualities, competencies and (overall) score (range)	Duration (in minutes)
Personal qualities	not applicable	Semi-structured Interview (SSI) • Motivation (1-3) • Orientation on the job (1-3) • Learning needs (1-3) • Personal attributes (1-3) Overall score (1-3)[a]	30 – 45
Knowledge	'knows'	National General Practice Knowledge Test (LHK) Score (0-100%)	150
Cognitions on how to deal effectively with practical dilemmas	'knows how'	Situational Judgement Test (SJT) Score (0-100%)	60
Self-reported competencies	'shows how'	Patterned Behaviour Descriptive Interview (PBDI) • Empathy (1-5) • Collaborative skills (1-5) • Coping with pressure (1-5) • Respect (1-5) • Self-care (1-5) • Reflection (1-5) Overall score (1-5)[b]	60
General clinical competencies	'shows how'	Simulated encounter (SIM) • Medical expertise (1-5) • Doctor-patient communication (1-5) • Professionalism (1-5) Overall score (1-5)[c]	30

[a] *average of the scores of the 4 qualities*
[b] *average of the scores of the 6 competencies*
[c] *average of the scores of the 3 competencies*

Table 2 Individual characteristics; difference, (δ, 95% confidence interval, CI) between study population and remaining candidates

Domains	Study population *n* = 47[a]	Remaining candidates *n* = 76	Difference δ (95% CI)	All *n* = 123
Gender, % male	29.8	25.0	4.8 (-11.5 – 21.1)	26.8
Age in years, mean (SD)	28.4 (2.8)	29.4 (5.2)	-1.0 (-1.9 – 0.9)	29.0 (4.5)
Past performance, % < 1 year clinical experience	38.3	36.8	1.5 (-16.2 – 19.1)	37.4
Region of medical school, % NW Europe	100	88.2	11.8 (4.6 – 19.1)	92.7
Times of application, % first time	83.0	78.9	4.4 (-13.1 – 21. 9)	80.5
Admitted, %	78.7	73.7	5.0 (-10.3 – 20.4)	75.6

[a]Nijmegen n = 23; Utrecht n = 24

Table 3 Mean and overall scores (SD) of the instruments in the current (SSI) and the new procedure (LHK, SJT, PBDI, SIM) (*n* = 47)

Instruments (n items)	Min – max	Mean (SD)	Number with poor assessments, score '1' or below threshold (%)
Semi-structured Interview SSI overall (4)	1.1 – 2,8	2,2 (0.3)	Not applicable
Motivation[a]	1.0 – 3.0	2,2 (0.4)	4 (8.7%)
Orientation on the job[a]	1.0 – 3.0	2,1 (0.4)	5 (10.9%)
Learning needs[a]	1.0 – 3.0	2,2 (0.4)	3 (6.5%)
Personal attitude[a]	1.0 – 3.0	2,1 (0,5)	5 (10.9%)
National General Practice Knowledge Test[a] (114)	6.7 – 43.3	21.9 (8.7)	10 (21.7%)
Situational Judgement Test (80)	75.1 – 91.0	83.8 (3.1)	6 (12.8%)
Patterned Behaviour Descriptive Interview overall(6)	1.3 – 4.7	3.7 (0.5)	6 (12.8%)
Empathy	1 – 5	3.8 (0.9)	1 (2.1%)
Collaboration	1 – 5	3.7 (0.8)	1 (2.1%)
Coping with pressure[a]	1 – 5	3.8 (0.8)	1 (2.2%)
Respect[b]	1 – 5	3.6 (0.8)	2 (4.7%)
Self-care[c]	1 – 5	3.5 (0.9)	2 (4.4%)
Reflection[d]	2 – 5	3.6 (0.7)	0
Simulates encounter overall (3)	1.3 – 4.7	2.9 (0.6)	12 (25.5%)
Medical expertise	1 – 4	2.4 (0.8)	7 (14.9%)
Doctor patient communication	2 – 5	3.1 (0.7)	0
Professionalism	1 – 5	3.1 (0.7)	1 (2.1%)

[a] *n=46*
[b] *n=43*
[c] *n=45*
[d] *n=44*

Table 4 Associations between scores of the SSI (current procedure), the PBDI en the SIM (new procedure); Pearson's correlation coefficient r (*n* = 47)

		SSI				**PBDI**						**SIM**		
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
SSI	Motivation	1												
	Orientation on the job	0.43[b]	1											
	Learning needs	0.62[b]	0.38[a]	1										
	Personal attitude	0.51[b]	0.24	0.70[b]	1									
PBDI	Empathy	0.00	0.26	-0.02	-0.01	1								
	Collaboration	-0.07	-0.15	-0.14	-0.07	0.48[b]	1							
	Coping with pressure	0.05	-0.04	0.21	0.10	0.31	0.50[a]	1						
	Respect	-0.07	-0.11	-0.25	-0.26	0.47[a]	0.35[a]	0.35[a]	1					
	Self-care	0.18	0.12	0.17	0.06	0.24	0.24	0.26	0.41[b]	1				
	Reflection	0.14	0.00	-0.15	-0.07	0.30[a]	0.29	0.00	0.15	0.19	1			
SIM	Medical expertise	0.25	0.17	0.25	0.33[a]	0.27	0.19	0.25	0.23	0.07	-0.06	1		
	Communication	0.08	0.02	0.31[a]	0.25	0.28	0.17	0.44[b]	0.01	-0.02	-0.25	0.32[a]	1	
	Professionalism	0.18	0.03	0.30[a]	0.32[a]	0.22	0.07	0.27	-0.03	0.11	-0.01	0.40[b]	0.76[b]	1

SSI: Semi-structured Interview; PBDI: Patterned Behaviour Descriptive Interview; SIM: Simulated Encounter
[a]p< 0.05 [b]p< 0.001

Table 5 Association between individual characteristics and (overall) scores of the instruments of the current and the new procedure; Pearson's correlation coefficient r (*n* = 47)

	Individual characteristics				Current	New procedure			
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Gender	1								
Age	-0.21	1							
Past performance	0.19	0.18	1						
Times of application	-0.08	0.10	0.33[a]	1					
Semi-structured Interview (*n*=46)	0.36[a]	-0.19	0.28	-0.07	1				
National GP Knowledge Test (*n*=46)	0.19	-0.25	0.08	-0.25	0.14	1			
Situational Judgement Test	0.21	-0.22	0.11	-0.21	0.14	0.34[a]	1		
Patterned Behaviour Descriptive Interview	-0.17	0.25	0.21	0.10	-0.03	-0.12	0.06	1	
Simulated Encounter	0.29	0.19	0.17	-0.11	0.34[a]	0.07	0.20	0.25	1

[a] *p< 0.05*

Table 6 Candidates' evaluation scores of the instruments of the current and new procedure on 5 point scale, ranging from 1 (strongly disagree) to 5 (strongly agree) *n* = 46

	Current	**New procedure**			
Domains	**SSI**	**LHK *n* = 45**	**SJT**	**PBDI**	**SIM**
Content instrument is relevant for post-graduate GP training mean (SD)	4.2 (0.6)	4.5 (0.7)[a]	4.2 (0.7)	4.2 (0.6)	3.7 (0.8)[a]
Instrument enables to show competence for postgraduate GP training mean (SD)	3.3 (0.8)	2.8 (0.9)[a]	3.5 (0.9)	3.9 (0.8)[a]	3.6 (0.7)
Instrument is fair to candidates % yes	63%	64%	96%[a]	74%	78%[a]

SSI: Semi-structured interview; LHK: National General Practice Knowledge Test; SJT: Situational Judgement Test; PBDI: Patterned Behaviour Descriptive Interview; SIM: Simulated encounter [a]*$p < 0.05$*

Chapter 8

General discussion

The studies described in the current thesis were conducted between 2009 and 2013 using clerical data collected from the administrative processes of the Dutch postgraduate GP training at both the national and local levels in Utrecht and Nijmegen. The studies can be considered part of the quality assurance cycle of the Dutch GP training. The findings and insights derived from this cycle were, at the same time, limited by it, as no other data were collected than those that were routinely available. In this chapter, we reflect on these findings within the existing literature and on current and future issues related to practice and research in the field.

Reliability and fairness of the semi-structured interviews

In 2005, the postgraduate GP training changed from a profession-based training to a competency-based training.[1] Until 2014, the selection procedure, however, remained profession-based, as it relied on the semi-structured interview.[2] The interviews were conducted by a committee consisting of a staff member, a trainer and a trainee. Reduction from three to two assessors hardly diminished the reliability of the semi-structured interview, especially when the trainee did not participate.[3] The most obvious reason for this finding might be that trainees can participate as interviewers only a limited number of times, whereas the other assessors might have more experience. The literature indicates that reliability will be improved with increased structuring of the interview and by conducting several short interviews.[4-6]

Because the department of choice influenced the chance of being admitted, we doubted the fairness of the procedure for the candidates.[2] This finding served as the basis of an argument in favour of a national selection procedure. To reduce the impact of local cultural influences, we endeavoured the implementation of national workshops for assessors and the interchange of candidates and assessors among the different GP departments. The ratio between the number of candidates and the number of vacancies differs among departments and is primarily geographically determined.[7] This rural issue is comparable to that of other countries, and it might result in a loss of suitable candidates who could become competent GPs.[8-10] The fact that not all available vacancies are filled and that GPs are geographically stable in choosing a practice near their GP training area, led to the decision of the HON to introduce a national allocation system in 2015.[11-13] The effect of this intervention is unknown, as trainees only want to participate in the training in unpopular regions if they can choose the training area voluntarily.[12] This finding provides an argument for thoroughly monitoring the national allocation system.

Poor performance and attrition

Poor performance is a common problem in all postgraduate medical training programs.[14-21]

Our study found that most poor performers and their trainers appeared capable of solving their shortcomings, which may be related to their reflective skills and active learning competencies. A small group, however, had more serious, repeated or 'chronic' problems. From earlier studies, we know that poor performers have difficulties in recognising their shortcomings; for example, they are more likely to over-rate themselves and to ask non-clinical assessors for assessments.[22-24] Consequently, these poor performers may also be at risk for poor performance as practising professionals.[17,25-28] Although we explored the competency areas in which the shortcomings of the poor performers occurred, we do not know whether the problems originated from burn out, personality problems, life events or other factors. This knowledge could be helpful in adapting remediation strategies to the needs of the poor performer and could, therefore, lead to increased success.[29,30] If remediation is not successful, involuntary attrition is sometimes the eventual solution.

In our study, we did not focus on voluntary attrition, in which trainees' attitudes play a prominent role.[14-16,31-37] Voluntary attrition results in the inefficient use of the training program and may be due to the trainee's inadequate understanding of the specialty. Thorough knowledge of the GP profession might reduce false expectations and would require accurate information from the GP context. Interventions might help. For example, Longitudinal Integrated Clerkships have demonstrated success in improving students' attitudinal and professional development.[38,39] With respect to surgery, Kelz et al. found that screening candidates by requiring them to write an essay related to organisational skills dramatically reduced attrition.[40]

Risk factors

The risk factors for poor performance identified within our cohort may have consequences for the selection procedure and the training program.

Age. During the selection process, it is necessary to explore the reasons why the candidate is older than other candidates. For example, did he/she drop out of another postgraduate training program, and if so, why? Was there stagnation during the undergraduate training, and if so, why? When an older trainee is admitted, staff members and trainers could work with the trainee to determine whether he/she is prone to or at risk of performing poorly. For example, is the candidate's field-specific knowledge out-dated? Does he/she have family responsibilities that could affect the balance between training and personal life? How much resilience does the candidate have with regard to engaging in new clinical situations and educational settings? If necessary, can he/she easily overcome out-dated knowledge and skills? The results of this exploration may indicate that older trainees require additional guidance and supervision.

Early knowledge and competencies assessments. The fact that early insufficient assessments are a risk factor for poor performance pleads for including early assessments in a programmatic assessment approach.[20,41,42] In fact, the selection procedure should be considered as an assessment at time zero and should be recorded in the assessment program. The procedure must contain instruments that assess knowledge and competencies congruently with the competencies to be acquired during the training program.[43] The results of the selection may be used to create an Individualised Development Plan.[44] The earlier poor performance is identified, the better the chance of the trainee successfully completing the training will be.[15,28,45-47] Trainees should learn how to solve their problems themselves, as this is an essential competence necessary for life-long learning.[48]

Poor performance in the previous year. This risk factor justifies close monitoring of poor performers, as they are at risk of showing poor performance in the current year. In general, in medical education examination procedures, it is recommended that assessments are conducted without prior knowledge to prevent bias. However, in our GP program, where workplace-based assessments together with more objective assessments are combined for formative and summative objectives, we emphasise that shortcomings in competencies are identified early. A comparison with medicine can be made in that - as doctors always approach ‘risky patients’ differently - staff members and trainers should collect information on the trainees’ previous performances and monitor poor performers more strictly. This approach will not only allow the staff members and trainers to focus on the shortcomings of the poor performers but also allow them to observe the coping ability of the trainee. Coping ability includes reflective skills and assuming responsibility for one’s own learning process, which are skills that lead to improved professional behaviour and clinical competences.[49,50]

Consistent with these findings, it seems appropriate, in the selection procedure, to have knowledge of the applicant’s performance during undergraduate training to enable medical students at risk of academic failure to be identified and supported.[46,51] Furthermore, the format of a competency-based assessment program might be analogous to a children’s screening program of growth and development. In other words, once the indicators for progress are satisfactory, the frequency of screening might decrease. However, if progress development is unclear or disturbed, more observations might be needed until enough information is obtained to identify possible poor performers.[52]

In addition, we advocate more training and support for trainers and staff members of postgraduate training programs to address the needs of poor performers for several reasons. First, because trainers do not always quantitatively indicate shortcomings of trainees in their workplace based assessment scores.[20,30] Second, because of the ‘fail to fail problem’, which means that staff members and trainers fail to fail the poorly performing student or trainee when faced with a student or trainee who is underachieving.[29,53,54] Finally, because remediation strategies are not always tailored to the needs of the poor performer, the strategies may not lead to an improved chance of success.[46,55,56]

Competency based selection procedure

The newly developed selection procedure that consists of the national GP knowledge test (LHK), a situational judgement test (SJT), a patterned behaviour descriptive interview (PBDI) and mini-simulations (SIM) has been adapted to the competency-based training and has been aligned with the current assessment principles.[43,57] The strength of this procedure resides in the fact that these instruments assess different levels of Millers pyramid ('knows', 'knows how', 'shows how') and that all information is aggregated to substantiate high-stake decisions regarding admission.[58]

The Dutch National postgraduate GP training (Huisartsopleiding Nederland, HON) decided to abandon the SIM on financial grounds and the SJT because of its poor psychometric properties. The cost of the revised selection procedure using the four instruments was estimated to be four times higher than that of the semi-structured interview. Although the SIM requires the greatest financial resources, it also significantly contributes to improved reliability and validity, and together with the LHK and PBDI, these instruments provide the best opportunities for obtaining valuable feedback.[58] Thus, the advantages may well outweigh the costs. First, the feedback can be used when establishing an Individualised Development Plan for the trainee at the start of training; second, problems may be identified earlier; and third, candidates will better understand, before the start of the program, what will be expected of them during the training. In addition, we can doubt whether this comprehensive selection procedure is necessary for all candidates.

Current and future issues

In the Netherlands, the project 'de Arts van Straks' ('The physician of tomorrow') was initiated in 2001 with the aim of shortening the time required to complete medical education. The rationale for this change was that once a student starts working as a specialist, he/she is 'too old, too clever and too expensive'.[59,60] Thus, the recommendation was to change the last year of the master's program into a 'transition year'. In this year, the student would function as a 'semi-doctor' and work under supervision but would have more responsibilities than a student during his/her clerkship. Since 2004, all Dutch medical schools have provided a 'transition year', which is intended to abbreviate the time needed to orient on a postgraduate training. A second recommendation was to make the entry requirements per specialty more transparent and to connect the undergraduate training seamlessly to the postgraduate training program. In 2014, 10% of the trainees are to be recruited for postgraduate training by this route, and the goal is to increase this percentage in the coming years.

The 'transition year' can enable a successful transfer to a postgraduate training program when the learning outcomes of undergraduate training are tightly connected to the entry conditions for postgraduate training. The selection procedures for postgraduate training may suffice as a marginal assessment of specific competencies per specialty, whereas differentiation in admission may be needed for those medical doctors who do not directly enter a postgraduate training program. For this group, a more comprehensive selection procedure might be appropriate. At the same time, the importance of continuing one's medical education after postgraduate training should be emphasized; the 'continuing medical education' includes reregistration and accredited education but could also include assessments of doctors' performance, such as 'patient assessments' and 'peer assessments'.[61-64]

In short, it is time to define the 'Continuum of Medical Education' that includes undergraduate, postgraduate and continued medical education.[65,66] A well-designed portfolio that grows from undergraduate to postgraduate training and beyond for the practising professional might be a beneficial instrument.[67]

Future research

The selection procedure

In the near future, the quality of the new selection procedure will be evaluated; what is the reliability and predictive value of the LHK and the two PBDIs with two assessors each, which have been implemented nationally since 2014 by the HON. The outstanding results of the SJT in the UK and Belgium – the high predictive validity on job performance, the incremental validity on both cognitive tests and personality tests and the cost-efficiency – support the decision to a further development and implementation of our SJT, which presents methodological problems, probably due to the length of the SJT and the format used.[68-72]

In addition, the value of the grade point average* (GPA) in the Master phase of the training must be further investigated, as several studies have shown its predictive validity.[10,43,73-77] The GPA was implemented in 2007 in the Netherlands and is included on the international diploma supplement of the medical degree.

Poor performance and attrition

We plead for a national database of trainees' selection assessments, early assessments and training process characteristics in order to explore the association between these characteristics on the one hand and outcome measures as quality of competencies, voluntary and involuntary attrition, illness and burn out during the training on the other hand. In addition, we propose conducting qualitative research on the group of poor

**Grade Point Average: the average obtained by dividing the total number of grade points earned by the total number of credits attempted.*

performers to improve remediation strategies, by exploring the origins of their problems and their needs for remediation, such as reflective practice.[49,50] The exploration of what GP trainers and staff members need in order to address the 'fail-to-fail problem' and what types of interventions may help the trainees who are struggling to meet standards (e.g., training for assessments, reflective practice, tailored coaching) may be the next step in developing effective interventions.

Lastly, we plead for collaboration with the Netherlands Institute for Health Services Research (NIVEL, 'Huisartsenregistratie-project') for an extension of their national database with characteristics of GPs and their practices with personal characteristics, training program characteristics and outcomes as a practising GP in order to explore the association between individual and training characteristics and outcomes as a practising GP, such as attrition, burn out, practice policies and job satisfaction.[78-81]

References

1. Project Vernieuwing Huisartsopleiding. http://www.huisartsopleiding.nl/content.asp?bid=10028668&kid=10023284&fid=-1
2. Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, Tromp F, Graaf van der Y, Pieters HM. Selection for Dutch postgraduate GP training; time for improvement. Eur J Gen Pract 2012;18:201-5.
3. Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, Graaf van der Y, Damoiseaux RAMJ. Dutch postgraduate GP selection procedure; reliability of interview assessments. BMC Family Practice 2013, 14:43 doi:10.1186/1471-2296-14-43.
4. Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR: Predictive validity of the multiple mini-interview for selecting medical trainees. Med Educ 2009,43:767–775.
5. Donnon T, Paolucci EO. A generalizability study of the medical judgment vignettes interview to assess students' noncognitive attributes for medical school. BMC Med Educ 2008;8:58. doi:10.1186/1472-6920-8-58.
6. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ 2011;45:1181-9.
7. Tromp F, Mokkink HGA, Thoonen BPA, Bottema BJAM. Ongepubliceerde enquête: 'Redenen voor de keuze waar aios hun opleiding tot huisarts volgen.' VOHA Nijmegen 2009.
8. Clark TR, Freedman SB, Croft AJ, Dalton HE, Luscombe GL, Brown AM, Tiller DJ, Frommer MS. Medical graduates becoming rural doctors: rural background versus extended rural placement. Med J Aust 2013;199:779-82.
9. Avery DM, Wheat JR, Leeper JD, McKnight JT, Ballard BG, Chen J. Admissions factors predicting family medicine specialty choice: a literature review and exploratory study among students in rural medical scholars program. J Rural Health 2012;28:128-36.
10. Roberts C, Al-Alwan I, Prideaux D, Tekian A. Developing the science of selection into healthcare professions and specialty training within Saudi Arabia and the Gulf region. J Health Spec 2013;1:71-7.
11. Schoots MJ, Honkoop PJ, Dunselman JHAM, Joziasse IC. Waar wil 'de nieuwe huisarts' zich vestigen? Huisarts Wet 2012;55:494-9.
12. Hingstman L, Velden. Huisartsen honkvast bij keuze locatie. Huisarts Wet 2010;53:73.
13. Muurling JA, Damoiseaux RAMJ. Uniforme selectie en spreiding opleidingsplaatsen. Huisarts Wet 2014;57:18-9.
14. Roback HB, Crowder MK. Psychiatry resident dismissal. A national survey of training programs. Am J Psychiatry 1989;146:96-8.
15. Bergen PC, Littlefield JH, O'Keefe GE, Rege RV, Antony TA, Kim LT ea. Identification of high risk residents. J Surg Res. 2000;92:239-44.
16. Yaghoubian A, Galante J, Kaji A, Reeves M, Melcher M, Salim A, Dolich M, de Virgilio C. General Surgery Resident remediation and attrition. Arch Surg 2012;147:829-33.
17. Williams RG, Roberts NK, Schwind CJ, Dunnington GL. The nature of general surgery resident performance problems. Surgery 2009;145:651-8.
18. Zbieranowski I, Takahashi SG, Verma S, Spadafora SM. Remediation of residents in difficulty: a retrospective 10-year review of the experience of a postgraduate board of examiners. Acad Med 2013;88:1-6.
19. Reamy BV; Harman JH. Residents in trouble: An in-depth assessment of the 25- year experience of a single family medicine residency. Fam Med 2006;38:252-7.
20. Mitchell C, Bhat S, Herbert A, Baker P. Workplace based assessments of junior doctors: do scores predict training difficulties? Med Educ 2011;45:1190-8.
21. ABIM: "The problem resident." VHS videocassette produced by American Board of Internal medicine; Portland, Ore 1992.
22. Mitchell C, Bhat S, Herbert A, Baker Paul. Work-based assessments in Foundation Programme: do trainees in difficulty use them differently? Med Educ 2013;47:292-300.
23. Gow KW. Self-evaluation: how well do surgery residents judge performance on a rotation? Am J Surg 2013;205:557-62.
24. Lipsett PA, Harris I, Downing S. Resident self-other assessor agreement: influence of assessor, competency and performance level. Arch Surg 2011;146:901-6.
25. Papadakis MA, Hodgson CS, Teherani A, Kothatsu ND. Unprofessional behaviour in medical school is associated with subsequent disciplinary action by a state medical board. Acad Med 2004;79:244-9.

26. Yates J, James J. Risk factors at medical school for subsequent professional misconduct: multi-centre retrospective case-control study. BMJ 2010;340:c2040.
27. Resnick AS, Mullen JL, Kaiser LR, Morris JB. Patterns and predictions of resident misbehaviour- a 10-year retrospective look. Curr Surg 2006;63:418-25.
28. Yao DC, Wright SM. National survey of internal medicine residency program directors regarding problem residents. JAMA 2000;284:1099-104.
29. Cleland J, Leggett, H, Sandars J, Costa MJ, Patel R, Moffat M. The remediation challenge: theoretical and methodological insights from a systematic review. Med Educ 2013;47:242-51.
30. Schwind CJ, Williams RG, Boehler ML, Dunnington GL. Do individuals attendings' post-rotation performance ratings detect residents' clinical performance deficiencies. Acad Med 2004;79:453-7.
31. Laufenburg HF, Turkal NW, Baumgardner DJ. Resident attrition form family practice residencies: United States versus international medical graduates. Fam Med 1994;26:614-7.
32. Yeo H, Bucholz E, Ann Sosa J, Curry L, Lewis FR, Jones AT, Viola K, Lin Z, Bell RH. A national study of attrition in general Surgery training : which residents leave and where do they go? Ann Surg 2010;252:529-34.
33. Longo WE, Seashore J, Duffy A, Udelsman R. Attrition of categoric general surgery residents: results of a 20-year audit. Am J Surg. 2009;197:774-8.
34. Sullivan MC, Yeo H, Roman SA, Ciarleglio MM, Cong X, Bell RH jr, Ann Sosa J. Surgical residency and attrition: defining the individual and programmatic factors predictive of trainee losses. J Am Coll Surg,2013;216:461-71.
35. Dodson TF, Webb ALB Why do residents leave general surgery? The hidden problem in today's program. Curr Surg 2005;62:128-31.
36. Katzenbauer M. 'Die cultuur paste niet bij mij.' Med Cont. 2009;64:580-3.
37. Kuyvenhoven MM, Vermeulen MI, van Campen SM, Schmidt JET. Veel aiossen vallen uit. Med Cont. 2010;65:2206-7.
38. Myhre DL, Woloschuk W, Pedersen J.S. Exposure and attitudes toward interprofessional teams: a three-year prospective study of longitudinal integrated clerkship versus rotation-based clerkship students. J Interprof Care. 2013 Sep 3. [Epub ahead of print] doi:10.3109/13561820.2013.829425.
39. Teherani A, Irby DM, Loeser H. Outcomes of different clerkship models: longitudinal integrated, hybrid, and block. Acad Med. 2013;88:35-43.
40. Kelz RR, Mullen JL, Kaiser LR, Pray LA, Shea GP, Drebin JA, Wirtalla CJ, Morris J. Prevention of surgical attrition by a novel selection strategy. Ann Surg 2010;252:537-43.
41. van der Vleuten CPM, Schuwirth LW: Assessing professional competence: from methods to programmes. Med Educ 2005,39:309–17.
42. Driessen E, F.Scheele. What is wrong with assessment in postgraduate training? Lessons from clinical practice and educational Research. Med Teacher 2013;35:569-74.
43. Prideaux D, Roberts C, Eva K , Centeno A, McCrorie P, McManus C et al. Assessment for selection for the health care professions and specialty training: Consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach 2011;33:215-23.
44. Challis M. AMEE Medical Education Guide No. 19: Personal learning plans. Med Teach 2000;22:225-36.
45. Evans DE, Alstead EM, Brown J. Applying your clinical skills to students and trainees in academic difficulty. Clin Teach 2010;7:230-5.
46. Steinert Y. The "problem" learner: Whose problem is it? AMEE Guide No. 76. Med Teach. 2013;35:e1035-e1045.
47. Artino AR Jr, Hemmer PA, Durning SJ, Using self-regulated learning theory to understand the beliefs, emotions and behaviours of struggling medical students. Acad Med 2011;86 (10 Suppl):35-8.
48. Taylor DCM, Handy H. Adult learning theories: Implications for learning and teaching in medical education: AMEE Guide No. 83 2013; 35:e1561–e1572.
49. Mann K, Gordon J, Macleod A. Reflection and reflective practice in health professions education: a systematic review. Adv Health Sci Educ Theory Pract 2009;14:595-621.
50. Kilminster S, Cottrell D, Grant J, Jolly B. AMEE Guide No. 27: Effective educational and clinical supervision. Med Teach. 2007;29:2-19.

51. Stegers Jager KM. At Risk Medical Students. Characteristics and possible interventions [dissertation]. Rotterdam 2012.
52. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. Med Educ 2012:46:38-48.
53. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. Acad Med 2005; 80 (10 suppl):S84-S87.
54. Roberts NK, Williams RG. Hidden costs of failing to fail residents. J Grad Med Educ. 2011;3:127-9.
55. Wu JS, Siewert B, Boiselle PM. Resident evaluation and remediation: a comprehensive approach, J Grad Med Educ 2010;2:242-5.
56. Scott Smith C, Stevens NG, Servis M. A general framework for approaching residents in difficulty. Fam med 2007;39;331-6.
57. Dijkstra J, Van der Vleuten C, Schuwirth L. A new framework for designing programmes of assessment. Adv Health Sci. Educ. 2010;15:379-93.
58. Vermeulen MI, Tromp F , Zuithoff NPA, Pieters HM, Damoiseaux RAMJ, Kuyvenhoven MM. A competency based selection procedure for Dutch postgraduate GP training: A pilot study on validity and reliability. Eur J Gen Pract 2014; Early Online: 1–7.
59. Meyboom de Jong B, Schmit Jongbloed LJ, Willemsen MC. 'De Arts van straks'. Utrecht: KNMG/DMW/VSNU/VAZ/NVZ/LCVV, 2002.
60. Heineman MJ, Borleffs JCC, Broek WW van den, Graaf J. de. Het schakeljaar uit de mottenballen. Med Cont 2014;69:442-4.
61. Overeem K, Faber MJ, Arah OA, Elwyn G, Lombarts KMJMH, Wollersheim HC, Grol RPTM. Doctor performance assessment in daily practice: does it help doctors or not? A systematic review. Med Educ 2007:41:1039-49.
62. Hays RB, Davies HA, Beard JD, Galdon LJM, Farmer EA, Finucance PM, McCroirie P, Newble DI, Schuwirth LWT, Sibbald GR. Selecting performance assessment methods for experienced physicians. Med Educ 2002;36:910-7.
63. Norcini JJ. Current perspectives in assessment: the assessment of performance at work. Med Educ 2005;39:880-9.
64. van Mook WN, Gorter SL, De Grave WS, van Luijk SJ, Wass V, Zwaveling JH, Schuwirth LW, Van Der Vleuten CP. Bad apples spoil the barrel: Addressing unprofessional behaviour. Med Teach. 2010;32:891-8.
65. Royal College of Physicians and Surgeons of Canada. The Continuum of Medical Education. 2011.
66. Association of American Medical Colleges. Teaching for Quality. Integrating Quality Improvement and Patient Safety Across the Continuum of Medical Education Report of an Expert Panel. 2013.
67. Van Tartwijk J, Driessen EW. Portfolios for assessment and learning: AMEE Guide no.45. Med Teach 2009;31:790-801.
68. Lievens F, Buyse T, Sackett PR. The operational validity of a video-based situational judgment test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains. J Appl Psychol 2005;90:442-52.
69. Lievens F, Patterson F. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. J Appl Psychol 2011;96:927-40.
70. Patterson F, Lievens F, Kerrin M, Zibarras L, Carette B. Designing selection systems for medicine: the importance of balancing predictive and political validity in high stakes selection contexts. Int J Select Assess 2012; 20;486-96.
71. Patterson F, Ashworth V, Zibarras L, Coan P, Kerrin M, O'Neill P. Evaluations of situational judgement tests to assess non-academic attributes in selection. Med Educ 2012;46:850-68.
72. Lievens F. Adjusting medical school admission: assessing interpersonal skills using situational judgement tests. Med Educ 2013:47:182-9.
73. Siu E, Reiter HI. Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. Adv in Health Sci Educ 2009;14:758-75.
74. Greenburg DL, Durning SJ, Cohen DL, Cruess D, Jackson DL. Identifying medical students likely to exhibit poor professionalism and knowledge during internship. J Gen Intern Med 2007;22:1711-7.
75. Kulatunga-Moruzi, Norman GR. Validity of admission measures in predicting performance outcomes: the contributions of cognitive and non-cognitive dimensions. Teach Learn Med. 2002;14:34-42.
76. Poole P, Shulruf B. Rudland J, Wilkinson T. Comparison of UMAT scores and GPA in prediction of performance in medical school: a national study. Med Educ 2012;46:163-71.
77. Cohen-Schotanus J, Muijtjens AM, Reinders JJ, Agsteribbe J, van Rossum HJ, van der Vleuten CP. The

predictive validity of grade point average scores in partial lottery medical school admission system. Med Educ. 2006;40:1012-9.

78. Hingstman, L, Kenens, R.J. Cijfers uit de registratie van huisartsen: peiling 2011. Utrecht, NIVEL, 2011.

79. Berg MJ van den, Kolthof ED, Bakker DH de, Zee J van der. Tweede Nationale Studie naar ziekten en verrichtingen in de huisartspraktijk: de werkbelasting van huisartsen. Utrecht, NIVEL, 2004.

80. Schmit Jongbloed LJ, Schonrock-Adema J, Borleffs JCC, Stewart RE Cohen-Schotanus J. The influence of achievements before, during and after medical school on physician job satisfaction. Acad Med 2014; Jan 24. [Epub ahead of print]

81. Houkes I, Winants Y, Twellaar M, Verdonk P. Development of burnout over time and the causal order of the three dimensions of burnout among male and female GPs. A three-wave panel study. BMC Public Health 2011;11:240. doi:10.1186/1471-2458-11-240

Summary

In *Chapter 1*, the aims of the thesis are defined. The first aim concerns reliability aspects of the Dutch selection procedure as carried out till 2014; the second aim concentrates on the frequency and nature of poor performance and attrition; the third aim relates to the construction and first execution of a new competency based selection procedure.

Chapter 2. A semi-structured interview has been the core of the Dutch selection procedure for postgraduate GP training until 2014. A staff member, a trainer and a trainee independently assess personal qualities in all eight department. We investigated to what degree department of choice, candidates' characteristics and qualities assessed during interviews explain admission into GP training in a nationwide observational study of all candidates who applied for postgraduate GP training in 2009/2010 (n = 597). The study population addressed 542 candidates. Sixty three candidates were rejected on letter of application (11.6 %). So 479 candidates were admitted to the interview, of which 340 were admitted to the GP training (71.0 %). Male candidates and candidates who followed medical school outside north western Europe had more risk of being rejected on letter of application. Department of choice had a strong association with admission in both stages (RR 0.30 – 0.74; 0.20 – 0.79, respectively) while candidates' qualities explained admission as well (RR 1.09 – 1.25). The influence of department of choice yields doubts about fairness and public defensibility of the decentralized Dutch selection procedure.

Chapter 3. Aiming to improve the selection procedure, we investigated the inter-rater reliability of the interview for groups of two or three assessors in an observational study of all candidates who entered the Utrecht selection procedure between April 2008 and 2010 (n = 394). Twenty-six candidates were rejected based on their application letter. Ultimately, 206 of the 365 candidates were admitted to the GP training. The inter-rater reliability was satisfactory (ICC 0.78 – 0.84). Reduction from three to two assessors slightly reduced the reliability. The candidates' qualities independently explained admission to the GP training (RR 1.34 – 1.84). Individual characteristics were not associated with the admission decision. These results did not differ for candidates who applied for the first time versus candidates applying for the second or third time. We concluded that selection interviews with two assessors yielded a satisfactory level of reliability.

Chapter 4 and 5. In an observational cohort study of trainees who started the GP training in Utrecht between 2005 and 2007 (n = 215), we investigated the frequency, nature and risk factors of poor performance and attrition among trainees in the first year of training, Chapter 4, and during the three years, Chapter 5. Trainees exhibited poor performance in 9.7%; 13.0% and 7.4% in the respective years. Overall percentage during the training was 22.8%. Six trainees were dismissed (2.8%). In the 1st and 2nd years, problem areas among poor performers were distributed equally across the roles as 'medical expert', ‘communicator’ and ‘professional’. In the 3rd year, a shortcoming in ‘professionalism’ was the most common problem. Higher age was a risk factor for poor performance; OR 1.16 (CI 1.06 – 1.27). Trainees with early sufficient assessment scores in ‘communication’ and knowledge were at lower risk of poor performance; OR 0.50 (CI 0.33 – 0.77) and OR 0.16 (CI 0.07 – 0.40), respectively. Poor performance in the previous year was a risk factor for poor performance in the 2nd and 3rd years; OR 4.20 (CI 1.31 – 13.47) and OR 5.40 (CI 1.58 – 18.47), respectively. Conclusion of Chapter 5 is that poor performance is prevalent, primarily occurring within a single training year. This result suggests that trainees and their trainers mostly are capable of solving trainee problems. In light of these findings, we recommend early assessments of competencies and additional monitoring and guidance of trainees who are ‘at risk’.

Chapter 6. The development process of a competency based selection procedure for the Dutch postgraduate GP training is described. We advocated an assessment procedure with multiple sources of information from various methods to construct an overall judgement by triangulating information across these sources. First, the content of the procedure was determined with a modified Delphi procedure. Then, we selected instruments to be used in the procedure by searching the literature for relevant, feasible, reliable and valid methods. Consensus on the following CanMEDS‘ roles was reached: ‘medical expert’, ‘communicator’, ‘collaborator, ‘manager’, and ‘professional’.
Four instruments were included: the National GP Knowledge Test (LHK); a Situational Judgement Test (SJT); a patterned behaviour descriptive interview (PBDI), and a series of three work-related simulations (SIM). The results of the selection instruments can be considered as feedback for the candidates and used at the start of the training for future development.

Chapter 7. In a pilot study of 47 candidates reliability and validity aspects of the new procedure have been explored. The procedure is feasible and standardisation is possible. Content-related competencies showed low correlations with one another when measured with different instruments, whereas more diverse competencies measured by a single instrument showed strong to moderate correlations. The LHK and SJT were easy to implement, but had poor intraclass correlation (ICC 0.59; 0.55, respectively) while the ICC of PBDI and SIM showed acceptable levels of reliability (ICC 0.79; 0.73, respectively), but were more challenging in implementation. Perception of the candidates was that the new instruments provided the same or better opportunities for demonstrating GP competence and were more fair to candidates.

Chapter 8 reflects on the main findings of the thesis and on current and future issues in practice and research. The result that two assessors hardly reduced reliability within a semi-structured interview in case of three assessors, has been incorporated in the new developed selection procedure. The finding that the department of choice influenced the chance of being admitted was an argument to introduce a national selection procedure by the Dutch National postgraduate GP training (Huisartsopleiding Nederland, HON) for 2014.

Monitoring poor performance from the beginning of the GP training is fruitful as poor performance in the previous year is a strong risk factor for exhibiting poor performance later on. Poor performers need extra guidance and support. Early assessments (including selection assessments) should be recorded in the assessment program. We pleaded for a national database of trainees' selection assessments, early assessments and training process characteristics to enable further research of risk factors for poor performance among trainees.

The newly developed competency based selection procedure consisting of the LHK, a SJT, a PBDI and SIM assesses candidates on different levels of Miller's pyramid. The assessments are suitable as formative feedback and can be used in the Individualised Learning Plan. The HON decided to a national selection procedure with LHK and PBDI in 2014 without SIM and SJT, due to costs and poor psychometric properties. Considering the costs, we argued for a more differentiated selection: use of the assessments attributed in the 'transition year' for the recruitment of trainees who choose for a postgraduate training directly after their MD graduation and to administer the more comprehensive procedure for the remaining candidates.

Next we recommended studies with regard to the validity and reliability of the new competency based procedure, further development of the SJT, and the prognostic value of the GPA with regard to competencies as assessed during the training program. Qualitative research in the group of poor performers on the origin of the problems and their needs to solve them, could give insight in possible effective remediation interventions. In addition exploring the needs of the trainers and staff members who are engaged with trainee assessment might help in addressing 'the fail to fail' problem.

Lastly we pleaded for collaboration with the NIVEL 'Huisartsenregistratie-project' for an extension of their national database with characteristics of GPs and their practices with personal characteristics, training programme characteristics and outcomes as a practising GP in order to explore the association between individual and training characteristics and outcomes as a practising GP within varying situational contexts.

Samenvatting

In *Hoofdstuk 1* worden de doelen van het proefschrift beschreven. Het eerste doel was het nagaan van de betrouwbaarheid van de Nederlandse selectie procedure, zoals deze werd uitgevoerd tot 2014; het tweede doel de exploratie van de frequentie en de aard van poor performance en uitval in de Utrechtse Huisartsopleiding het laatste doel betrof de ontwikkeling en een eerste pilot van een nieuwe competentie gerichte selectie procedure.

Hoofdstuk 2. Tot 2014 was een semi-gestructureerd interview het belangrijkste onderdeel van de selectie procedure voor de huisartsopleiding in Nederland. Op alle 8 huisartsinstituten beoordeelden een staflid, een opleider en een aios persoonlijke kwaliteiten van de kandidaten. We onderzochten in welke mate het instituut van keuze, kenmerken en kwaliteiten van de kandidaat de toelating tot de huisartsopleiding verklaarden. Hiertoe werd een landelijke observationele studie uitgevoerd naar alle kandidaten die solliciteerden voor de huisartsopleiding in 2009/2010 (*n* = 597). De studiepopulatie bestond uit 542 kandidaten. Drieënzestig kandidaten werden afgewezen op basis van hun sollicitatie brief (11,6%) en 479 kandidaten werden toegelaten tot het interview; hiervan werden er 340 aangenomen tot de huisartsopleiding (71,0%). Manlijke kandidaten en kandidaten die de geneeskunde opleiding buiten Noord West Europa volgden, hadden meer kans om te worden afgewezen op hun sollicitatie brief. Het instituut van keuze was sterk geassocieerd met toelating tot beide fasen van de selectie (respectievelijk RR: 0,30 – 0,74; 0,20 – 0,79). De kwaliteiten van de kandidaten verklaarden eveneens de toelating tot de huisartsopleiding (RR 1,09 – 1,25). Het feit dat het instituut van keuze zoveel invloed had op de kans om toegelaten te worden leverde twijfel op over de eerlijkheid en de verdedigbaarheid van de decentrale selectieprocedure in Nederland.

Hoofdstuk 3. We onderzochten de interbeoordelaars betrouwbaarheid van het interview voor groepen van drie of beoordelaars in een observationele studie van alle kandidaten die aan de Utrechtse selectieprocedure participeerden tussen april 2008 en 2010 (*n* = 394). Zesentwintig kandidaten werden op brief afgewezen. Uiteindelijk werden 206 van de 365 kandidaten aangenomen voor de huisartsopleiding. De interbeoordelaars betrouwbaarheid was voldoende (ICC 0,78 – 0,84). Vermindering van drie naar twee assessoren verlaagde de betrouwbaarheid nauwelijks. De kwaliteiten van de kandidaten verklaarden de toelating tot de huisartsopleiding (RR 1,34 – 1,84). Persoonlijke kenmerken als sekse en leeftijd hingen niet samen met de beslissing een kandidaat toe te laten. Deze resultaten verschilden niet tussen kandidaten die voor het eerst solliciteerden en kandidaten die voor een tweede of derde keer solliciteerden. De conclusie was dat met twee assessoren een voldoende mate van betrouwbaarheid wordt behaald.

Hoofdstuk 4 en 5. In een observationele cohort studie van aios die begonnen aan de huisartsopleiding in Utrecht tussen 2005 en 2007 (*n* = 215), onderzochten we de frequentie en aard van, en de risicofactoren voor poor performance en uitval onder aios in hun eerste jaar (hoofdstuk 4) en gedurende de hele opleiding (hoofdstuk 5). Overall was er sprake van poor performance bij 22,8 % van de aios; 9,7% in het eerste jaar, 13,0% in het tweede en 7,4% in het derde jaar. Zes aios werden uit de opleiding gezet. In het eerste en tweede jaar waren de probleem gebieden van deze groep gelijkelijk verdeeld over de competentiegebieden 'vakinhoudelijk handelen', 'communicatie' en 'professionaliteit'. In het 3e jaar bleken de lacunes mn op het gebied van 'professionaliteit' te liggen. Oudere leeftijd was een risicofactor voor poor performance; OR 1,16 (CI 1,06 – 1,27). Aios die in het eerste kwartaal voldoende beoordeling hadden op 'communicatie' en kennis hadden een lagere kans op poor performance; respectievelijk OR 0,50 (CI 0,33 – 0,77) en OR 0,16 (CI 0,07 – 0,40). Poor performance in het voorgaande jaar bleek een risicofactor voor poor performance in het tweede en derde jaar; respectievelijk OR 4,20 (CI 1,31 – 13,47) en OR 5,40 (CI 1,58 – 18,47). Conclusie van hoofdstuk 5 was dat poor performance regelmatig voorkomt, meestal gedurende één enkel jaar. De resultaten suggereren dat aios en opleiders veelal in staat zijn om de problemen op te lossen door lacunes weg te werken. Gezien deze bevindingen is onze aanbeveling om vroege beoordelingen van kennis en competenties tijdens de opleiding te handhaven en de risico aios extra te monitoren en te begeleiden.

Hoofdstuk 6. Hierin wordt het ontwikkelingsproces van een competentiegerichte selectie voor de huisartsopleiding in Nederland beschreven. We pleitten voor een beoordelingsprocedure met meerdere informatiebronnen en verschillende methodes om tot een uitspraak te komen over de geschiktheid van de kandidaat om aan de opleiding te beginnen. Als eerste is de inhoud van de procedure bepaald met behulp van een gemodificeerde Delphi procedure uitgaande van het competentieprofiel van de huisarts. Daarna hebben we de instrumenten gekozen op basis van een literatuurstudie door te zoeken naar relevante, haalbare, betrouwbare en valide methodes. Er werd consensus bereikt in de volgende CanMEDS taakgebieden: 'vakinhoudelijk handelen', 'communicatie', 'samenwerken', 'organiseren' en 'professionaliteit'. Vier instrumenten werden geïncludeerd: de Landelijke Huisartsgeneeskundige Kennistoets (LHK), een Situational Judgement Test (SJT); een gestructureerd interview (PBDI) en een serie van 3 werk gerelateerde simulaties (SIM). De resultaten van de selectie instrumenten kunnen dienen als feedback voor de kandidaten en kunnen gebruikt worden voor verder ontwikkeling aan het begin van de opleiding.

Hoofdstuk 7. In een pilotstudie bij 47 kandidaten zijn betrouwbaarheids- en validiteitsaspecten van de nieuwe procedure geëxploreerd. De procedure is haalbaar en standaardisatie is mogelijk. Inhoud gerelateerde competenties correleerden zwak met elkaar indien ze gemeten werden met verschillende instrumenten, terwijl meer inhoudelijk verschillende competenties gemeten met één instrument een matige tot sterke correlatie liet zien. De LHK en de SJT konden makkelijk geïmplementeerd worden, maar hadden een lage intra class correlaties (respectievelijk ICC 0,59; 0,55). De ICCs van de PBCI en SIM toonden een acceptabele mate van betrouwbaarheid (ICC respectievelijk 0,79; 0,73). De kandidaten vonden dat de nieuwe instrumenten dezelfde of betere mogelijkheden gaven om competenties aan te tonen. Tevens vonden ze de nieuwe instrumenten rechtvaardiger.

Hoofdstuk 8. Het laatste hoofdstuk reflecteert op de belangrijkste bevindingen van het proefschrift en richt zich op huidige en toekomstige kwesties ten aanzien van onderzoek en praktijk. Het resultaat dat met twee assessoren de betrouwbaarheid van een semi-gestructureerd interview nauwelijks minder wordt, is al geïncorporeerd in de nieuw ontwikkelde selectie procedure. De bevinding dat het instituut van keuze de kans om aangenomen sterk beïnvloedde was een belangrijk argument voor Huisartsopleiding Nederland (HON), om een landelijke selectie procedure te implementeren met ingang van 2014.

Het monitoren van poor performance vanaf het begin van de huisartsopleiding is zinvol omdat poor performance in het voorafgaande jaar een sterke risicofactor is om opnieuw onder te presteren. Deze aios hebben extra begeleiding en ondersteuning nodig. Vroege beoordelingen (waaronder die van de selectie procedure) van competenties en kennis moeten worden opgenomen in het toets programma. We pleiten voor een landelijke database met daarin de beoordelingen van aios tijdens de selectie, vroege beoordelingen, voortgang en opleiding gerelateerde kenmerken voor verdere studie naar risicofactoren voor onvoldoende voortgang en problemen van aios .

De nieuw ontwikkelde competentie gerichte selectie procedure, bestaande uit de LHK, een SJT, een PBDI en SIM, beoordeelt de kandidaten op verschillende niveaus van de piramide van Miller. De beoordelingen zijn geschikt als formatieve feedback en kunnen gebruikt worden in Individuele Opleidings Plannen. De HON heeft besloten tot een landelijke procedure met de LHK en PBDI in 2014, maar zonder de SIM en SJT vanwege respectievelijk de kosten en de matige psychometrische eigenschappen. De kosten meewegend, pleiten wij voor een gedifferentieerde selectie: gebruik de beoordelingen tijdens het '(dedicated) schakeljaar' voor werving en selectie van aios die een medische vervolgopleiding ambiëren direct na het behalen van het artsexamen en gebruik een uitgebreidere procedure voor de overige kandidaten, die niet direct doorstromen.

Vervolgens adviseren we in de nabije toekomst de betrouwbaarheid en validiteit van de nieuwe competentie gerichte procedure te onderzoeken, de SJT verder te ontwikkelen en de predictieve waarde van het Grade Point Average ten aanzien van de beoordeelde competenties tijdens de opleiding uit te zoeken. Kwalitatief onderzoek onder de poor performers om de oorsprong van hun problemen te exploreren en beeld te krijgen van wat

zij nodig hebben om problemen op te lossen, kan inzicht geven in effectievere remediëring strategieën. Daarbij zou exploratie van de behoeften van opleiders en stafleden die aios moeten beoordelen kunnen helpen om het 'fail to fail' probleem aan te pakken.
Tenslotte pleiten we voor samenwerking met het NIVEL 'Huisartsenregistratie-project' voor een uitbreiding van hun landelijke database met kenmerken van huisartsen en hun praktijken met persoonlijke en opleidings kenmerken en uitkomsten als praktiserend huisarts om samenhang hier tussen te exploreren.

Dankwoord

Dankwoord

Promoveren kan je niet alleen. Het is fantastisch om te ervaren dat er velen meer of minder betrokken zijn geweest bij dit jarenlange proces. Mijn dank hiervoor is groter dan ik duidelijk kan maken in dit dankwoord. Daarbij realiseer ik mij terdege dat ik een heel bijzonder traject heb mogen volgen waarin ik veel vrijheid heb genoten en mogelijkheden heb gekregen om mij verder te ontwikkelen.

Allereerst gaat mijn oprechte dank uit naar mijn promotoren en co-promotoren.
Beste Roger, in het begin was ik een beetje huiverig; wat heeft een clinicus en 'orenman' met onderzoek van onderwijs? Die angst was niet nodig; al tijdens je docentjaar bij de huisartsopleiding raakte je betrokken bij mijn onderzoek en in korte tijd heb jij je daar vol enthousiasme in verdiept. Je gaf mij vertrouwen dat ik dit traject zou kunnen afmaken. Ik kon altijd bij je terecht en voelde mij steeds door jou gesteund. En samen verkenden wij de laatste officiële promotiefase die voor ons beiden, vanuit een ander perspectief, nieuw was. Heel veel dank!
Beste Yolanda, erg zenuwachtig was ik voor onze eerste afspraak in de Bilt waarin ik voorzichtig mijn plannen voorlegde. Tenslotte waren dit geen doorsnee onderzoeksplannen zoals men gewend was binnen het Julius. Wat knapte ik op van het feit dat jij er wel wat in zag; je gaf mijn onderzoek hiermee bestaansrecht. En dat ben je blijven doen. We zagen elkaar niet vaak, maar tijdens de gesprekken die we voerden, ervaarde ik je rake opmerkingen vanuit verschillende invalshoeken als zeer waardevol! Dank hiervoor!
Beste Marijke, zonder jou had dit boekje er echt nooit gelegen! Je bent bij elke fase volop betrokken geweest, hebt steeds meegedacht en mij op sleeptouw genomen. Hierdoor werd ik mij langzaam bewust dat onderzoek van onderwijs een bijzondere tak van sport is. Jouw achtergrond, je gedrevenheid en je precisie hebben mij enorm geholpen. Ik vind het geweldig dat je na je pensioen de begeleiding wilde continueren. Jouw arbeidsethos is bewonderenswaardig, je interesse voor mij als persoon was heel plezierig. Ik mis onze maandagmiddagen! Dank voor alles!
Beste Ron, jij hebt een hele cruciale rol gespeeld in dit proces. Ik weet nog steeds niet hoe je het voor elkaar hebt gekregen. Tenslotte zette ik jaren geleden op papier dat ik echt niet van plan was te gaan promoveren en als je dat wel van mij verwachtte, je beter iemand anders kon zoeken voor de managementfunctie. Maar jij had geduld en stuurde bijna onmerkbaar. Je gaf mij veel ruimte om mijn eigen vragen te mogen beantwoorden en je gaf mij vertrouwen dat ik het zou kunnen. Ik ben je daar erg dankbaar voor.

Prof. dr. Edith ter Braak, Prof. dr. Ronald Bleys, Prof. dr. Jan van Schaik, Prof. dr. Matthijs Numans, Prof. dr. Jan Borleffs, beste leden van de beoordelingscommissie. Hartelijk dank voor het beoordelen van mijn proefschrift.

Beste Peter: met engelengeduld heb je me steeds weer statistische en methodologische zaken uitgelegd zonder me daardoor heel erg dom te laten voelen. Gelukkig nam mijn eigen kennis tijdens het traject toe en kon ik op een gegeven moment meedenken en zelf analyses doen. Met jouw kennis en humor voerden wij discussies over wat wel en niet kon. Je eerdere betrokkenheid bij de SVUH gaf een extra dimensie aan onze samenwerking en ik hoop dat we in de toekomst samen projecten kunnen blijven doen!

Beste Luc, jij bent heel belangrijk voor mij geweest in de eerste jaren als staflid van de huisartsopleiding vooral op het gebied van ontwikkeling (en niet alleen van onderwijsprogramma's). Al vroeg wekte jij mijn onderzoeksinteresse en legde daar een basis. Beste Raf, vers en onwennig in het managementteam leerde jij mij hoe ik me mijn nieuwe rol kon invullen en gaf je me ruimte om met onderzoek te beginnen en de master Epidemiologie te volgen. Beste Rien, onze gesprekken gingen over ontwikkeling en onderzoek in de breedste zin van het woord. Ik wil jullie alle drie hier hartelijk voor bedanken.
Beste Ben, Hans, Harry, Fred. Ineens mocht ik met jullie mee op werkbezoek in Oxford om te zien hoe de aios huisartsgeneeskunde werden gerekruteerd. Dank! En jammer Jan dat je op het laatste moment niet meekon!
Beste Hans en Shanna, dank voor het uitzoeken van de uitval uit de huisartsopleiding en meeschrijven aan het artikel in Medisch Contact. Beste Kees, heel veel dank voor de interesse en investering vanuit de SBOH in de selectiepilot.
Beste Jan, Nettie, Herman, Paula en Angela, onderzoek doen is leuk, maar uitvoeren in de praktijk is echter en spannender. Dank voor deze mogelijkheid en de prettige samenwerking!
Dank ook aan de werkgroep toetsing, in het bijzonder Paul, Anneke, Harry en Marjan, van wie ik veel heb kunnen leren, hetgeen ik goed kon gebruiken bij de ontwikkeling van de nieuwe selectieprocedure.
En bij die ontwikkeling en uitvoering horen Mariëlle en Edger. Wat verfrissend om met andere professionals te werken en te discussiëren. Een verrijking, waar ik enorm van genoten heb! Dank voor jullie hulp en steun. Voor mij zijn jullie meer dan collega's geworden. En speciale dank aan Barend voor zijn hulp bij de ontwikkeling en uitvoering van de SJT.
Alle medewerkers van de huisartsopleiding Utrecht; dank voor jullie belangstelling, gezelligheid en relativerende woorden. Ik realiseer me dat ik niet altijd zichtbaar ben geweest en zie er naar uit om daar verandering in te brengen. Dank aan Loes, Tecla, Carolien, Marianne, Jacqueline en in het bijzonder Monique: jullie hebben het mogelijk gemaakt dat ik mij kon focussen op het onderzoek door me te ondersteunen en taken over te nemen; ik heb dit enorm gewaardeerd. Beste O2ers en Ivan, ik zie er naar uit om veel

meer met jullie te gaan werken. En dank aan Charles, Rianne, Bert Jan, Anne en Dorien voor pep talks op momenten dat het heel erg welkom was. Dank ook aan het co team op de vrijdag voor de gezelligheid.

Gezelligheid was er ook buiten het werk, dank aan vrienden, familie en kennissen die mij afleidden en terug wierpen op kleine en grote dingen in het leven tijdens borrels, feestjes, etentjes, voorstellingen, wandelingen en tijdens het samen muziek maken.

En dan mijn paranimfen: Fred, toevallig kruisten onze wegen elkaar en we hadden niet kunnen vermoeden dat we zoveel samen zouden gaan mee maken in de afgelopen jaren. Soms zat het mee en soms zat het tegen. Wat bijzonder dat ik je heb mogen leren kennen en ik hoop dat we contact houden! En Mirjam, zonder jou was het allemaal een heel stuk lastiger geweest. Jij nam zonder enig probleem veel van mij over. En geen moment heb ik me hoeven afvragen of het allemaal wel goed liep. Je warme belangstelling, je nuchterheid, je verbazing en de spiegel die je me voorhield, voelden als een cadeautje. Dank!

Lieve Lily, je hield me steeds in de gaten door met oprechte belangstelling te informeren naar waar ik mee bezig was, dat deed me goed. Lieve Robert, Floor, Kiki en natuurlijk Annemarie, jullie gaven de nodige afleiding en dat is heel belangrijk voor mij geweest.
Lieve papa en mama; tja zonder jullie was ik er niet geweest en dankzij jullie heb ik kunnen worden wie ik ben. Dank voor die basis en het geloof en vertrouwen in mij. Jullie zagen al veel eerder dan ikzelf dat er een onderzoeker in mij schuilde.
Lieve Ilse en Veerle, wat een voorrecht om jullie moeder te zijn, te zien hoe jullie je ontwikkelen en daar zelf weer van te leren. En het goeie (?) nieuws is; ik heb nu al meer tijd om me nog meer met jullie te bemoeien!
Lieve Gaston, al zolang mijn lief, mijn maatje, mijn steun. Jij, die altijd al in mij geloofde. Jou daarvoor danken is niet te beschrijven. Wij raken niet uitgepraat over het rijke leven. Met jou is het leven mooier!

Curriculum Vitae

Curriculum Vitae

Margit Ilse Vermeulen was born on September 21st 1965 in Eindhoven. After graduating from secondary school (gymnasium, St Joris College, Eindhoven) in 1983, she studied medicine at Utrecht University. From 1991, she worked as a MD in hospitals in Harderwijk, Rotterdam and Gouda at departments of surgery, internal medicine, intensive care and burn centre.

She completed the postgraduate GP training from 1995 - 1998. She worked as a GP in several practices in Utrecht and surroundings. She practiced with Brenda Ott (GP) for several years in Zeist through 2008.

From 1998 she was appointed at the postgraduate GP training Utrecht as a staff member especially for the development of educational material. Later on she tutored groups of GP trainees in their first year. In 2008 she became manager Research and Development of the postgraduate GP training and joined the national assessment group of Huisartsopleiding Nederland (HON).

In 2009 she started her PhD studies by investigating her own questions arisen in the educational environment she has been working. In 2012 she obtained her Master of Science degree in Clinical Epidemiology at Utrecht University, with a minor at the School of Health Education in Maastricht.

Currently she is working as manager Research and Development at the Utrecht GP Department (UMC Utrecht) and she has been partly seconded at the HON on behalf of the national selection procedure. Her main focus in research is improvement of the selection procedure for the postgraduate GP training.

Margit Vermeulen is married to Gaston van de Laar. They have two daughters, Ilse and Veerle.

Publications

Vermeulen MI, Tromp F, Zuithoff NP, Pieters HM, Damoiseaux RA, Kuyvenhoven MM. A competency based selection procedure for Dutch postgraduate GP training: A pilot study on validity and reliability. Eur J Gen Pract 2014 Mar 19. doi:10.3109/13814788.2014.885013

Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, Van der Graaf Y, Damoiseaux RAMJ. Hoe goed is de selectie voor de huisartsenopleiding? HuisartsWet 2014;57:10-3.

Vermeulen MI, Kuyvenhoven MM, Zuithoff NP, van der Graaf Y, Damoiseaux RA. Dutch postgraduate GP selection procedure; reliability of interview assessments. BMC Fam Pract 2013 14:43. doi: 10.1186/1471-2296-14-43.

Vermeulen MI, Kuyvenhoven MM, Zuithoff NP, Tromp F, van der Graaf Y, Pieters HM. Selection for Dutch postgraduate GP training; time for improvement. Eur J Gen Pract 2012; 18:201-5. doi:10.3109/13814788.2012.680588

Zwart DLM, Heddema WS, Vermeulen MI, Van Rensen EL, Verheij TJ, Kalkman CJ. Het melden van incidenten door huisartsen in opleiding. Huisarts Wet 2012;55:200-3.

Zwart DLM, Vermeulen MI. Patiëntveiligheid in de opleidingspraktijk. Bijblijven 2012;28: 43-7.

Zwart DL, Heddema WS, Vermeulen MI, van Rensen EL, Verheij TJ, Kalkman CJ. Lessons learnt from incidents reported by postgraduate trainees in Dutch general practice. A prospective cohort study. BMJ Qual Saf 2011;20:857-62.

Vermeulen MI, Kuyvenhoven MM, Zuithoff NP, van der Graaf Y, Pieters HM. Attrition and poor performance in general practice training: age, competence and knowledge play a role.
Ned Tijdschr Geneeskd 2011;155:A2780. Dutch.

Kuyvenhoven MM, Vermeulen MI, van Campen SM, Schmidt JE. Veel aiossen haken af. Med Cont 2010;65:2206-7.

Van Geel AN, Hazelbag HM, Slingerland R, Vermeulen MI. Disseminating adamantinoma of the tibia. Sarcoma 1997;1:109-11.

Vermeulen MI, van Vroonhoven TJ, Leguit P. Acute appendicitis: a serious disease in the elderly. Ned Tijdschr Geneeskd. 199;139:1635-8. Dutch.

Berends FJ, Vermeulen MI, Leguit P. Perforation rate and diagnostic accuracy in acute appendicitis. Ned Tijdschr Geneeskd. 1994;138:350-4. Review. Dutch.