Functional assessments

A study on functional ability in a population, and structured assessments in general practice

Nina Østerås

Section for Social Medicine

Institute of General Practice and Community Medicine

University of Oslo, Norway

2009

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo
No. 834

ISBN 978-82-8072-965-1

Cover: Inger Sandved Anfinsen.
Printed in Norway: AiT e-dit AS, Oslo, 2009.

Produced in co-operation with Unipub AS.
The thesis is produced by Unipub AS merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

Unipub AS is owned by
The University Foundation for Student Life (SiO)

To Mum and Dad
for being who you are
and for your lifelong support

Table of contents

Acknowledgements

When I graduated as a physiotherapist twelve years ago, a doctoral education was not part of my ambitions. However, I am now very grateful for the series of coincidences, which provided an opportunity to take this education. There are many people that have supported and helped me during this process, but first of all I want to thank my two supervisors, Søren Brage and Pål Gulbrandsen, for their genuine involvement, their quick and clear responses, and for critically reading and commenting on numerous manuscript drafts. Your long, broad experience and your constructive, positive feedback were invaluable. Additionally, I thank my third supervisor, Gunnar Tellnes, for your support, genuine interest, and concern in my work.

This project work would not have been so nice and not so much fun without the project group members: Søren Brage, Kariann Krohne and Pål Steiran. Furthermore, I am very grateful for the contributions of my co-authors: Jurate S. Benth, Bård Natvig, Andrew Garratt, Fredrik Dahl, Dag Hofoss and Inger Cathrine Kann. Special thanks to Jurate for always being helpful and positive, and to Bård for the invitation to take part in the Ullensaker Study 2004. Further, many thanks to Therese Nordberg for her accurate and important work in the data collection period during my maternity leave. I also want to thank Ola Thune at NAV for his work with the register file, Magne Thoresen for statistical advices, and Cheryl Carling for translating the Norwegian Function Assessment Scale into English.

The present work has been given financial support by the Norwegian Ministry of Labour and Social Inclusion and has been carried out at the Section for Social Medicine, Institute of General Practice and Community Medicine, University of Oslo. I am very grateful for all valuable feedback I have received on manuscript drafts and for important discussions with my colleagues at the institute, and in particular with my colleagues at the section. The reference group for the project, “Functional Assessments - Health Providers’ Responsibilities in a More Inclusive Working Life”, have also given valuable feedback and comments to my work. Furthermore, without the participants in the pilot test, The Ullensaker Study 2004, and in the RCT in general practice, this work could not have been carried out.

Last but not least, many thanks to my husband, my parents, my brother, and my friends who have shown interest and have encouraged and supported me. I am also very grateful for my husband’s care, his flexibility and his invaluable computer skills, and for our two children, who daily give us lots of joy and essential breaks from our theoretical work.

Definitions and abbreviations

In this thesis the following definitions of core terms will be used:

Sickness certification	A declaration issued by a medical doctor, usually a general practitioner, to a person entitled to sickness benefits when this person is found to be incapacitated for work because of disease, illness or injury (1)
Sick-listed person	A person with medically certified absence from work due to disease, illness or injury
Sick leave	Medically certified absence from work due to disease, illness or injury (1)
Sick leave episode	The period of consecutive calendar days in which a person is declared by a medical doctor to be incapacitated for work
Sick leave case	An instance of consecutive sick leave attached to one person
Long-term sick leave	Sick leave episode lasting for more than eight weeks (1)
Part-time sick leave	A sick leave certification option allowing the employee to be absent from work for a specified proportion of the working hours or work week
Active sick leave	The Norwegian social insurance option that enables people on sick leave to attend work doing other tasks than they normally do. The National Insurance Administration provides 100% remuneration of normal wages during the active sick leave period for a maximum length of eight weeks.
Vocational rehabilitation	Support or allowance granted to a sick-listed person who need to change job or job training because of ill health
Self-efficacy	Beliefs in one’s capabilities to organize and execute the courses of action required to produce given attainments (2)

Abbreviations

CI	Confidence interval
COOP/ WONCA	The Dartmouth Primary Care Cooperative Information Project/ World Organization of National Colleges, Academies and Academic Associations of General Practitioners/Family Physicians
EQ-5D	EuroQol five dimensions
GP	General practitioner
ICC	Intraclass correlation coefficient
ICF	The World Health Organization's International Classification of Functioning, Disability and Health
NAV	The Norwegian Labour and Welfare Service
NFAS	The Norwegian Function Assessment Scale
NHP	Nottingham Health Profile
RCT	Randomised controlled trial
SF-36	The Short Form 36-item Health Survey
WAI	The Work Ability Index
WHO	World Health Organization

Functional assessments and functional ability assessments mean the same and will be used interchangeably in this thesis.

Summary

Functional assessments – A study on functional ability in a population, and structured functional assessments in general practice

Background

There is an increased focus on functional ability assessments in relation to sick leave. As a consequence, general practitioners (GP) in European countries are to an increasing extent being asked to assess function, in addition to disease and illness, in social security claims. For the GPs, this means paying attention to patient resources, possibilities, and coping, rather than symptoms, problems, and limitations. The GPs report difficulties in performing the requested explicit functional assessments. This could be due to lack of training and guidelines, as well as confusing terminology and insufficient knowledge of specific occupational demands. The Norwegian Function Assessment Scale (NFAS) is an instrument for self-report that was developed in 2000 by an expert group in social insurance to assess the need for rehabilitation, adjustment of work demands among sick-listed persons as well as the rights to social security benefits. For NFAS score interpretations, further validation and normative functional ability data were necessary.

The first aim of this study was to obtain population based normative data for the NFAS and further validate and test two versions of this instrument in a Norwegian population. The second aim was to implement structured functional assessments for persons with long-term sick leave in general practice in a cluster randomised controlled trial and assess intervention effects on important GP parameters, GP sick-listing practice, and patient sick leave.

Methods

The NFAS was included in a large, population based study, The Ullensaker Study 2004. All persons in seven birth cohorts in Ullensaker municipality in 2004 were approached by means of a postal questionnaire. Respondents were randomised to receive the original four- or the new five-point scale version of the NFAS. The results for the two versions were compared by evaluating data quality, internal consistency and validity. Functional ability scores by gender, age and education level were calculated, and the two-week test-retest reliability for the four-point scale version of the NFAS was assessed by total proportions of agreement, weighted kappa, and intraclass correlation coefficient (ICC).

In the cluster randomised controlled trial, 57 GPs were randomly assigned to an intervention or a control group. The invention GPs learned the structured functional assessment method at an introductory one-day workshop including teamwork and role playing, and they were requested to implement the method on ten consecutive sick-listed persons. The criteria for including a sick-listed person were: being part-time or full-time sick-listed for between eight and 26 weeks and having good prospects of a return to work. The intervention period ran from March 1st to October 31st in 2005. The outcome measures included GP knowledge, GP attitudes, and GP self-efficacy related to functional assessments, as well as GP knowledge of patients' work factors, and were collected before, immediately after and six months after the intervention period. Evaluation score-sheets were filled in by the intervention GPs and their patients immediately after the consultation. Sick leave data was extracted from the sick leave register of The Norwegian Labour and Welfare Administration.

Results

For the four-point scale version of the NFAS in The Ullensaker Study, the response rate was 54% (1620 persons). Items had low levels of missing values (3.3%). The test-retest response rate was 75% (101 persons), and the test-retest reliability was acceptable with high proportions of absolute agreement; kappa and ICC values ranged from 0.38 to 0.83 and 0.79 to 0.88, respectively. Thirty-three percent of the respondents reported no difficulty for all 39 functional activities. Females, older persons and persons with lower levels of education reported more functional problems than their respective counterparts ($p<0.05$). The age gradient was most evident for three of the physical domains. For females aged 24-56 and males aged 44-76, a clear education gradient was present for three of the physical domains and one mental domain after adjusting for age and gender.

Both the four- and the five-point scale versions of the NFAS had acceptable response rates and good data quality and internal consistency. The five-point scale version had somewhat better data quality in terms of missing data and end effects at the item and scale level. Furthermore, it had higher levels of internal consistency and item-discriminant validity. Construct validity was acceptable for both versions; demonstrated by correlations with instruments assessing similar aspects of health and comparisons with groups of individuals known to differ in their functioning according to existing evidence.

The intervention group GPs reported increased knowledge and self-efficacy related to functional assessments and increased knowledge about their patients' workplace and

perceived stressors. The intervention effects sustained at the second follow-up six months after the intervention period. There was no intervention effect on GP attitudes towards functional ability. Both before and after the intervention, the GPs were most informed about physical stressors, and less about mental and work organisational stressors. After the consultation, both the intervention GPs and their patients reported that the GPs' knowledge about the patients' work factors had increased.

The GP prescription of part-time sick leave was significantly higher ($p<0.01$) and the prescription of active sick leave was significantly lower ($p=0.04$) in the intervention group compared to the control group. There was no intervention effect on the duration of patient sick leave episodes or on GP prescription of vocational rehabilitation.

Conclusion

Population based normative data on functional ability, as measured by two scale versions of the NFAS, was collected in relation to gender, age and education level. The test-retest reliability of the four-point scale version and the validity for both versions were acceptable. However, the data quality, internal consistency and discriminative validity suggest that the five-point scale version of the NFAS should be used in future applications. The normative data is necessary for score interpretations and may serve as a basis for the development of national population norms.

The use of a structured functional assessment method in general practice led to significantly increased GP knowledge of functional assessments and work factors, and higher self-efficacy among the GPs with lasting effects at the second follow-up. It also changed the GPs' sick-listing practice by significantly increasing prescription of part-time sick leave and decreasing prescription of active sick leave. No intervention effect was seen on duration of patient sick leave episodes or on prescription of vocational rehabilitation.

Norsk sammendrag

Funksjonsvurderinger – En studie av funksjonsevne i en populasjon, og strukturerte funksjonsvurderinger i allmennpraksis

Bakgrunn

Det er for tiden et økende fokus på funksjonsvurderinger i forhold til sykefravær. På bakgrunn av dette, blir allmennleger i europeiske land stadig oftere bedt om å vurdere funksjonsevne i tillegg til sykdom og plager i sykmeldingssammenhenger. For legene medfører dette at de nå skal rette oppmerksomheten mot pasientressurser, muligheter og mestring i stedet for symptomer, problemer og begrensninger. Legene rapporterer imidlertid om problemer med å etterkomme kravet om en eksplisitt funksjonsrapportering. Dette kan skyldes manglende opplæring og retningslinjer, men også forvirrende terminologi og utilstrekkelig kjennskap til spesifikke arbeidskrav. Norsk Funksjonsskjema er et spørreskjema for selvrapportering av funksjonsevne. Skjemaet ble utviklet i år 2000 av en ekspertgruppe innen trygdemedisin og var tenkt brukt ved vurdinger av behov for rehabilitering og endring av arbeidskrav samt ved vurdinger av rett til varige ytelser. For å kunne tolke resultatene fra funksjonsvurderinger, trenger man å vite hva som er det gjennomsnittlige funksjonsnivået i en befolkning målt med et validert måleinstrument.

Det første målet med dette doktorgradsarbeidet var å undersøke funksjonsnivået i en populasjon ved hjelp av Norsk Funksjonsskjema og samtidig validere og teste to ulike versjoner av spørreskjemaet. Det andre målet var å implementere legebaserte strukturerte funksjonsvurderinger av langtidssykmeldte personer i en klinisk kontrollert studie i allmennpraksis og evaluere intervensjonseffekter i forhold til viktige legeparametre, legenes sykmeldingspraksis og pasientenes sykefravær.

Metode

Norsk Funksjonsskjema ble inkludert som en del av et spørreskjema i en stor befolkningsundersøkelse, Ullensakerundersøkelsen 2004. Alle personene i sju fødselskohorter i Ullensaker kommune i 2004 fikk tilsendt et spørreskjema i posten. Deltakerne ble randomisert til å motta en av to versjoner av Norsk Funksjonskjema, enten versjonen med fire svaralternativer eller versjonen med fem svaralternativer. Resultatene fra de to versjonene ble sammenlignet med hensyn til datakvalitet, intern konsistens og validitet. Gjennomsnittlige

funksjonsskårer for menn og kvinner, ulike alders- og utdanningsnivåer ble beregnet, og to-ukers test-retest reliabilitet ble analysert som totale andeler av enighet, vektet kappa og intraklasse korrelasjonskoeffisient.

I den randomiserte kontrollerte studien ble 57 allmennleger tilfeldig fordelt til en intervensjonsgruppe og en kontrollgruppe. Legene i intervensjonsgruppen fikk opplæring i en strukturert metode for funksjonsvurderinger og ble bedt om å bruke denne metoden på ti påfølgende sykmeldte personer. Inklusjonskriteriene for de sykmeldte personene var: helt eller delvis sykmeldt mellom åtte og 26 uker, samt ha gode utsikter til å kunne komme tilbake til arbeidet. Intervensjonsperioden varte fra 1. mars til 31. oktober i 2005. Utfallsmålene bestod av legenes kunnskap om funksjonsvurderinger, deres tiltro til egne mestringsevner (self-efficacy) og holdninger i forhold til funksjonsvurderinger i tillegg til kjenskap til pasientenes arbeidsfaktorer. Legene selv-rapporterte dette før, umiddelbart etter og seks måneder etter intervensjonsperioden. Evalueringsspørreskjemaer ble fylt ut av intervensjonslegene og deres pasienter umiddelbart etter konsultasjonen. Data om pasientenes sykefravær ble innhentet fra sykefraværsregisteret til NAV.

Resultater

I Ullensakerundersøkelsen var svarprosenten for Norsk Funksjonskjema med fire svaralternativer 54% (1620 personer). Det var et lavt nivå av ubesvarte spørsmål (3,3%). Svarprosenten for test-retestdelen var 75% (101 personer). Test-retest reliabiliteten var akseptabel med høy grad av enighet. Kappa og intraklasse korrelasjonskoeffisient-verdiene varierte fra henholdsvis 0,38 til 0,83 og 0,79 til 0,88. Trettitre prosent av deltakerne krysset av for kategorien “ingen vansker” på alle de 39 funksjonsaktivitetene i Norsk Funksjonskjema. Kvinner, eldre personer og personer med lavere utdanningsnivå rapporterte mer funksjonsproblemer enn deres respektive motstykker ($p<0,05$). Aldersgradienten var mest tydelig for tre av de fysiske funksjonsområdene. For kvinner i alderen 24-56 år og menn i alderen 44-76 år var det en tydelig utdanningsgradient for tre av de fysiske funksjonsområdene og ett av de mentale funksjonsområdene etter justering for alder og kjønn.

Begge versjonene av Norsk Funksjonsskjema hadde akseptabel svarprosent samt god datakvalitet og intern konsistens. Versjonen med fem svaralternativer hadde noe bedre datakvalitet i form av færre ubesvarte spørsmål og lavere tak- og gulveffekter for enkeltspørsmål og funksjonsområder. Videre hadde den bedre intern konsistens og diskriminerende validitet. Begrepsvaliditeten var akseptabel for begge versjonene;

demonstrert ved korrelasjoner med andre måleinstrumenter som måler lignende aspekter av helse samt ved sammenligninger mellom grupper av individer som i følge tidligere forskning er kjent for å ha ulikt funksjonsnivå.

Etter at intervensjonslegene hadde tatt i bruk den stukturerte metoden for funksjonsvurderinger, rapporterte disse legene om økt kunnskap om funksjonsvurderinger, økt tiltro til egne mestringsevner og økt kjennskap til deres pasienters arbeidsplasser og opplevde belastninger. Denne intervensjonseffekten var vedvarende ved den andre oppfølgingsrunden, seks måneder etter at intervensjonsperioden var avsluttet. Det var ingen signifikante endringer i legenes holdninger i forhold til funksjonsvurderinger som følge av intervensjonen. Både før og etter intervensjonsperioden var legene best informert om pasientenes fysiske belastninger, og mindre om mentale eller arbeidsorganisatoriske belastningsfaktorer. Etter konsultasjonen med funksjonsvurderingen, vurderte både legen og pasientene at legenes kjennskap til pasientens arbeidsplass og arbeidsoppgaver hadde økt.

Bruken av graderte sykepenger var signifikant høyere ($p<0,01$) og bruken av aktiv sykmelding signifikant lavere ($p=0,04$) blant intervensjonsgruppen enn blant kontrollgruppen. Det var en ingen intervensjonseffekt i forhold til pasientenes sykefraværslengde eller for bruken av yrkesrettet attføring

Konklusjon

Funksjonsnivået i en norsk populasjon, målt med to ulike versjoner av Norsk Funksjonsskjema, ble undersøkt i forhold til kjønn, alder og utdanningsnivå. Test-retest reliabiliteten av Norsk Funksjonsskjema med fire svaralternativer var akseptabel, og validiteten var god for begge versjonene. Resultatene i forhold til datakvalitet, intern konsistens og diskriminerende validitet tilsier imidlertid at versjonen med fem svaralternativer bør være den som benyttes i fremtiden. Dataene kan benyttes til å tolke funksjonsskårer, og de kan utgjøre en basis for utvikling av normaldata basert på den norske befolkningen.

Innføringen av en strukturert metode for funksjonsvurdering i allmennpraksis medførte at legene fikk økt kunnskap om funksjonsvurderinger og om ulike arbeidsfaktorer. Videre rapporterte legene høyere tiltro til egne mestringsevner i forhold til funksjonsvurderinger. Disse effektene var vedvarende ved den andre oppfølgingsrunden. Intervensjonslegene endret sin sykmeldingspraksis ved at de oftere brukte graderte sykepenger og sjeldnere aktive sykmeldinger. Det var ingen intervensjonseffekt på sykefraværslengden eller på bruken av yrkesrettet attføring.

List of papers

Paper I

Functional ability in a population: normative survey data and reliability for the ICF based Norwegian Function Assessment Scale.

Østerås N, Brage S, Garratt A, Benth JS, Natvig B, Gulbrandsen P.

BMC Public Health 2007, 7:278.

Paper II

A randomised comparison of a four- and a five-point scale version of the Norwegian Function Assessment Scale.

Østerås N, Gulbrandsen P, Garratt A, Benth JS, Natvig B, Brage S.

Health and Quality of Life Outcomes 2008, 6:14

Paper III

Implementing structured functional assessments in general practice for persons with long-term sick leave: a cluster randomised controlled trial.

Østerås N, Gulbrandsen P, Benth JS, Hofoss D, Brage S.

Accepted BMC Family Practice

Paper IV

Structured functional assessments in general practice increased the use of part-time sick leave: a cluster randomised controlled trial.

Østerås N, Gulbrandsen P, Kann IC, Brage S.

Resubmitted

1. Introduction

1.1. Historical background and terminology

1.1.1. The return of 'the functional approach'

The functional approach for understanding human illness used to be the norm until the end of the 18th century, when it was replaced by a biomedical model which emphasised clinical diagnostics (3). Although the biomedical model has generally proven valuable, it has over the past decades received criticism for being inefficient in the practice of general medicine and social insurance medicine (4;5). With an increasing proportion of long-term sick leave caused by musculoskeletal disorders, mental disorders, and subjective health complaints (6), the biomedical paradigm has proven inadequate in accounting for work disability following many of these conditions (5). In these cases, the assessment and diagnosis are seldom based on pathological changes, but rather on the general practitioners' (GP) interpretation of symptoms and the patients' subjective reports. Therefore, models based on a biopsychosocial approach have received increased attention in the recent decades (4). These models integrate the biomedical approach with a social understanding of illness, thereby emphasizing the individuals' ability to function within their environment. Thus, the functional approach is again of current interest and in use, not only in the practice of general medicine, but also in epidemiological research in general.

1.1.2. The international classification of functioning

The World Health Organization's International Classification of Functioning, Disability and Health (ICF), is an attempt to describe and classify functioning systematically (7). The ICF represents a revision of the former International Classification of Impairments, Disabilities and Handicap (ICIDH) that was released in 1980 by the World Health Organization (WHO). The ICF classification was approved by the WHO as a member of the WHO Family of International Classifications in 2001 (7). The main objective of the ICF is firstly to classify human health in relation to functioning and disability. It is meant to form a base for scientific studies of health and health related conditions, and provide a common language for transferring information between different participants, thereby allowing information exchange.

The ICF has a biopsychosocial understanding of functional ability, and the term functioning serves as the umbrella term for abilities encompassing four different components in the classification:

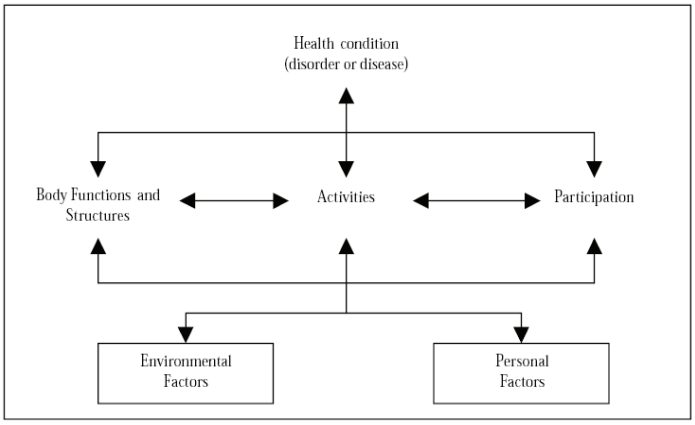
- Body functions : the physiological (and psychological) functions of body systems
- Body structures: the anatomical parts of the body
- Activity: the execution of a task or action by an individual person
- Participation: the involvement in life activities, a societal perspective of functioning

Contextual factors are the factors that together constitute the complete context of an individual person's life, and in particular the background against which health states are classified in the ICF. There are two components:

- Environmental factors: all aspects of the external or extrinsic world that form the context of an individual person's life and have an impact on that person's functioning. Environmental factors include the physical world and its features, the human made physical world, other people in different relationships and roles, attitudes and values, social systems and services, and policies, rules, and laws.
- Personal factors: contextual factors that relate to the individual person such as age, sex, social status, life experiences, etc.

The ICF is based on a model that integrates these components (see Figure 1-1).

Figure 1-1. ICF's model for functioning



As a classification tool, the ICF is complex and extensive, and short versions, core sets, for clinical and statistical purposes have been, and are being, made (8;9).

1.1.3. Core terms

Functioning

According to ICF, functioning is an overall term for all activities of human life. It refers to all body functions, activities and participation, and it represents the positive aspects of the interaction between a person (with a health condition) and the person's contextual factors (environmental and personal factors). The ICF and others (10) have pointed out the importance of relating the individual's functioning to a purpose and a context, e.g. in relation to workplace demands.

Disability

Disability is defined in the ICF as an umbrella term for impairments, activity limitations and participation restrictions (7). It represents the negative aspects of the interaction between a person (with a health condition) and the person's contextual factors (environmental and personal factors).

Functional ability

The functional ability describes the individual's capacity of functioning (7). According to the ICF, it represents a hypothetical capacity based on an assessment of the individual's potential and possibilities for action. The term, 'resources', and functional ability are often used synonymously. As with functioning, the functional ability must be related to the individual, the purpose, and the context. In medical and vocational rehabilitation, functional assessments represent a balancing of individual functional abilities against occupational demands and restrictions as a prerequisite for successful reintegration into working life.

The World Organization of Family Doctors (WONCA) defines functional ability as the ability of a person to cope with and adapt to changing elements in his or her individual environment, and to perform certain tasks to a measurable degree (11). According to this definition, functional ability contains different dimensions, e.g. work ability, physical fitness, and ability to cope with daily activities (12). In the present thesis, the ICF definition of functional ability as a hypothetical capacity will be used.

Work ability

Physical, mental and social capacities are important determinants for the individual to manage in the working life, and when these functional capacities are related to work demands, the term work ability can be used. However, work ability represents a complex issue reflecting individual and occupational factors that are essential for a person's ability to cope in the working life. Therefore, work ability also covers aspects like education, knowledge, skill, experience, and motivation (13). Further, the individual resources are influenced by the person's values, attitudes, and job satisfaction. In addition, the work ability is not only influenced by the physical and mental work demands, but also the work community, work environment, surrounding society, enterprise and the person's network (14).

As with functional ability, there exist different definitions of work ability. In a recent Norwegian official report the term work ability has been given a very wide meaning where the general functional ability is an integral part of the work ability (15). A more restricted definition is: 'Work ability is a person's physical and psychological capacity to perform his/her ordinary, remunerative work' (16). However, Ilmarinen and Tuomi have defined work ability as the ability of a worker to perform his or her job at present and in the near future, taking into account the specific work demands, individual health condition and mental resources (13). This definition will be used in the present thesis.

Functional ability versus work ability

Functional ability and work ability are strongly related terms, and since the distinction between them is not clear, they are sometimes used interchangeably. However, a few attempts have been done in order to do a distinction. In a recent Norwegian official report (15), the term functional assessments was mainly used in relation to GPs' evaluations of rights to social security benefits, whereas work ability assessments were connected to more total evaluations of a person's resources and limitations in relation to the demands and expectations from the working life. According to this view, a GP functional assessment cannot denote a complete work ability assessment including competence and professional experience, but rather represent a mapping of the patient's health condition and how this affect the patient's functional ability. This is in accordance with Schult's statement (17), that GPs can evaluate functional ability, whereas it is the employee and the employer that jointly should evaluate the employee's work ability and thereafter relate this ability to workplace adjustments.

1.2. Functional assessments and epidemiological research

1.2.1. Functional assessments may vary in relation to context and purpose

As with functional ability and work ability, there are different definitions and understandings of functional assessments. The various meanings or views often depend on the profession and the professional context (18). For some professions or situations, functional assessments may represent assessments of organ functioning, which could be related to specific disease conditions, whereas in other contexts it may represent assessments of the whole person as an integrated entirety.

There are a number of ways to classify functional assessment measurements. They can be classified according to what they are meant to describe: physical, mental and/or social functioning, or according to being general or body part specific. It is also possible to classify according to purpose, e.g. assessing work ability in sickness certification, rights to social security benefits or as a tool in relation to workplace adjustments and rehabilitation, or to evaluate effects of the adjustments/rehabilitation. Additionally, the measurements can be classified by whether they are self-administered, assessed by an expert or represent actual testing of ability or capacity (19).

An optimal scheme for assessing functional ability should be person orientated, clinically relevant, reliable and valid. Further, the scheme should be responsive for small changes (high responsiveness) and be easy to use. Additionally, an optimal scheme should be acceptable for all persons regardless of age, gender, ethnicity or cultural background. Finally, it should be internationally accepted as a standardised measure (20). However, instruments of proven reliability and validity in one country are often not properly translated, resulting in cross cultural problems (21).

Functional assessments should be kept distinct from pure functional descriptions as the functional assessments should include evaluations of what the person can do or might do in relation to the demands, expectations or requests that the person faces in his or her surroundings. Without this evaluation, no real assessment takes place. According to an American model (22) a GP functional assessment should consist of the following elements: the patient's own description of function, a discussion of possibilities and limitations, information about different demands at the workplace, and finally the GP's independent, total evaluation of medical and non-medical information.

The individual's own evaluation of his or her ability represents an important aspect, which can be mapped by a questionnaire or through a structured interview. A description of

functional ability based on what the person actually does can often be collected by self-reporting. There are other ways of collecting this, e.g. by observational or experimental studies. However, these are often difficult and/or costly methods (23).

1.2.2. Functional assessment instruments and item scaling

The measurement of functional ability is important in many contexts, e.g. in relation to obtaining population functional ability levels or for identifying necessary work demand or workplace adjustments. The most common approach for the measurement of functional ability is to ask individuals themselves, and a wide array of questionnaires, instruments and survey tools have been used in epidemiological and clinical studies. These instruments normally comprise a number of scales that measure different aspects of functioning, for example physical and role functioning. These scales, in turn, are comprised of a number of items or questions that relate to different aspects of functioning that are normally summed to produce a scale score.

The development of the content of questionnaires that measure functioning are normally based on some combination of a literature review, interviews with people who may or may not have an illness and expert consensus. During the construction of a function assessment scale, the selection of items and the grading of response categories are normally based on a consensus process among experts. Examination of the content of questionnaires reveals broad agreement in terms of instrument content including the different aspects of functioning covered.

There are a number of options open to developers of questionnaires including layout, ordering, presentation, and scaling of items or questions (24). While there often seems to be agreement as to the content of instruments for evaluation of function, there is relatively less consensus about the scaling of items. Item scaling vary in the number of response categories, the wording of category options and the use of all-point (where all categories are defined) or end-point (where only end-points are defined) scales (24;25). The majority of health status and patient-reported outcome measures use all-point defined scales with between two and seven categories, the most popular being five-point scales including the agree/disagree Likert format. The generic Short Form 36-item (SF-36) Health Survey (26;27) uses five-point scales for almost all of the scales it includes. Other generic instruments such as the Nottingham Health Profile (28) and EuroQol EQ-5D (29) use two- and three-point scales respectively. There is also considerable variation in the response scales for the large number of disease and

condition specific questionnaires with some developers opting for yes/no responses (30;31) while others have used seven all-point defined scales (32-34).

The optimal number of response categories is a matter that remains unresolved in spite of decades of research (35), with the debate dating back to a review in 1915 (36). Response alternatives should be so refined that they are capable of capturing most of the information available from respondents without being so refined that it encourages response error (37). Following a recent systematic review, it was recommended that future research designs should allocate respondents to different versions of a questionnaire to compare approaches to item scaling (24).

1.2.3. Short presentation of commonly used generic functional assessment instruments

The Short Form 36-item (SF-36) Health Survey

SF-36 is a generic measure with an eight-scale profile of functional health and well-being scores (26;27). The applications include general population surveys, clinical research, daily clinical practice and other areas including diverse populations (38). The eight scales are: Physical Functioning, Role-Physical, Bodily Pain, General Health, Vitality, Social Functioning, Role-Emotional and Mental Health. The eight scales are hypothesised to form two distinct higher ordered clusters, physical and mental health, due to the variance that they have in common. In the 2.0 version (SF-36v2) the Physical Functioning items have three response categories and one bodily pain item has six, whereas all other response choices are scored on a five-point category scale. For each question, the raw scores are coded, recalibrated in ten items, summed and transformed to the eight 0-100 scales (0=poorest possible health state, 100=best possible health state) according to the SF-36 scoring algorithms. The SF-36 is available in both standard (four-week) and acute (one-week) recall versions, except for the two scales, Physical Functioning and General Health, which do not have a recall period.

The SF-36 can be administered in five to ten minutes with a high degree of acceptability and data quality (27). According to Brazier et al. (39), it is easy to use, acceptable to patients, and fulfils stringent criteria of reliability and validity. The SF-36 has been judged to be the most widely evaluated generic patient assessed health outcome measure (40) and is suitable for self-administration, computerised administration, or administration by a trained

interviewer in person or by telephone, to persons aged 14 and older. The SF-36 is translated into 22 different languages.

The Nottingham Health Profile (NHP)

The NHP is a self-administered questionnaire developed to be used in epidemiological studies of health and disease (28). It is composed of two sections containing 45 items (41). The first section contains 38 items assessing six areas: sleep, physical mobility, energy, pain, emotional reactions and social isolation. Part 2 of the profile consists of seven statements relating to those areas of daily life most often affected by health: paid employment, jobs around the house, social life, personal relationships, sex life, hobbies and interests, and holidays. The two parts may be used independently. All items have a yes/no answer format. In part 1, positive answers are given the appropriate weighting, and the higher the score on any section, the greater the number and severity of perceived problems in that area. The maximum score on any section is 100. The NHP has undergone extensive evaluation and both strengths and weaknesses have been demonstrated (42).

EuroQol EQ-5D

The EQ-5D is a short generic instrument for describing and valuing health-related quality of life consisting of a self-classifier and a visual analogue scale (EQ-VAS) (29). In the current EQ-5D version (43), the respondents describe their own health state on five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. These are divided into three levels: no problems, some and severe problems. One of three levels are chosen for each dimension and this generates 243 different ‘health states’. Respondents also rate their perception of own overall health by means of the visual analogue scale ranging from 0 (worst imaginable health state) to 100 (best imaginable health state). This has the additional possibility of converting the descriptive data into values for economic (cost-effectiveness) analyses by linking patients' health state descriptions to empirical valuations of health states obtained from the general population. Algorithms have been developed for transforming health status information collected by the EQ-5D questionnaire into a single utility index value, ranging from −0.59 to 1.0, where 1.0 means optimal health in a population (44). Initially developed simultaneously in Dutch, English, Finnish, Norwegian and Swedish, EQ-5D is now available in most major languages and cultural adaptations (45). The EQ-5D is a well-established preference based health-related quality of life measure with acceptable feasibility, construct validity and discriminative ability (46)

The Dartmouth COOP Functional Health Assessment Charts/ WONCA (COOP/WONCA)

The COOP/WONCA is a generic health status measure, where functional status is self-reported with a time frame of the previous two weeks (21;47;48). The scale was developed for use in primary care settings and comprises six charts: Physical fitness, Feelings, Daily activities, Social activities, Overall health and Change in health. Each chart has five all-point defined response alternatives with pictorial representations with low scores indicating good functional ability. The instrument has demonstrated an acceptable distribution of the scores (49), good reliability (50-52) and validity (46;50;52), an acceptable responsiveness (50;51) and discriminative ability (46). It is easy to use and understand, feasible both in clinical practice and in research, and takes only three to five minutes to complete (50;53).

1.3. Functional assessments and social insurance

1.3.1. Short introduction to the Norwegian Social Insurance Scheme

All persons who are either residents or working as employees in Norway, are compulsorily insured under the National Insurance Scheme.

Sick leave: An insured person who has an annual income of at least 0.5 basic amount (NOK 35 128 at December 31, 2008) is entitled to daily cash benefits if he/she is incapable of working due to disease, illness or injury. It is, as a general rule, required that the person has been employed for at least four weeks. Daily cash benefits for employees equal to 100 per cent of pensionable income, and are paid from the first day of sick leave for a period of 260 working days (52 weeks). The employer pays the daily cash benefits for the first 16 calendar days, and thereafter the benefits are paid by the National Insurance Scheme. Income exceeding six basic amounts is not taken into account.

Vocational rehabilitation: For sick-listed persons who need to change job or job training because of ill health, vocational rehabilitation allowance is granted to insured persons between 19 and 67 years, whose ability to obtain employment income or possibility to choose occupation is permanently reduced by at least 50 per cent due to illness, injury or impairment. Furthermore, it must be considered necessary that the person goes through vocational training before he/she can get or keep suitable work. Full vocational rehabilitation allowance amounts to approximately 65% of previous income. The allowance is granted during waiting periods before and while the person is going through vocational training, and for up to three months after the vocational training are carried through while he/she applies for suitable work.

Active sick leave: Active sick leave is an option that enables people on sick leave to attend work doing other tasks than they normally do. The sick-listed person, the GP, the employer and the local social security officer may take the initiative, but initiation must be approved by the local social security office in advance, and the employee and the employer must make an activity plan. The National Insurance Administration provides daily cash benefits for employees equal to 100 per cent of pensionable income, during the active sick leave period. The maximum length for active sick leave is eight weeks. Exceptions can be made for this time limitation.

1.3.2. Functional assessments in social insurance

Starting in the UK in 1995 with the 'Personal Capability Assessment' (54), there is again an increasing interest for functional ability as a supplementary criterion to diagnosis in European social insurance. Development and further testing of structured functional assessment methods have been carried out in some countries, including England and Finland (55-57), and in some European countries, loss of functional ability has been introduced as an eligibility criterion for social security benefits (54;58;59). Among these, Denmark is the country that most prominent has put this trend into action by introducing the 'Arbejdsevnemetoden' in 2003 for assessing the need for disability pension among Danish workers.

From the middle of the 1990s there has been an increase in lost work days due to sick leave in Norway. To explore the reasons for this and to come up with ways to reduce this trend, the Norwegian Government appointed a committee, led by Sandman (60). Among different suggestions in the committee report, one was to emphasise assessments of functional ability: 'The functional assessment should indicate the functional ability/the functional disability in relation to the work and what must be done in order for the person on sick leave to return to his or her work' (own translation) ["Funksjonsvurderingen skal angi funksjonsevne/funksjonssvikt i forhold til arbeid og hva som skal til for at den sykmeldte skal komme tilbake i arbeid"]. Functional assessments were further described as an important tool for facilitating return to work and for assessing the need for social security benefits. Factors that were considered as important in this relation included the sick-listed persons' motivation, their ability to cope and their own assessment of the situation.

The 'Inclusive Working Life Agreement'

In line with the European trends, new ways to handle sick leave were introduced in Norway in October 2001 (61) as the Norwegian Government along with the employee unions and the employer associations agreed on a tripartite agreement ('Inclusive Working Life Agreement'). Through this agreement, several of the suggestions from the governmental report (60) were put into action. Hence, attention was placed on reducing sick leave and disability pensions, increasing the retirement age, and ensuring the recruitment of people with impaired functioning capacity and other vulnerable groups to the employment market. To achieve this, the focus was directed at functional ability, independently of disease and diagnosis, and work tasks should be individually adjusted. Employers and employees were explicitly given a larger responsibility to reintegrate the sick-listed persons by emphasizing a functional approach. The GPs should assist in this process by providing opinions on the sick-listed person's functional ability in relation to work demands. The inclusion of a Simplified Functional Assessment [Forenklet funksjonsvurdering] on sickness certification forms in 2002 represented an attempt to guide the assessment of this new functional criterion. The employers received copies of the GPs' comments on functional status on the certification forms, and were to use this information to provide workplace or workload adjustments.

Emphasizing work-related activities

New Norwegian rules for sickness certifications were implemented by July 1, 2004, which emphasised the importance of doing work-related activities during sick leave. A new sickness certification form, with an increased focus on resources and activity, was also introduced along with the new rules. Furthermore, the rules requested that the employer should make an action plan in order to facilitate a quick return to work for the sick-listed person. This plan should be made as soon as possible, but at the latest after eight weeks of sick leave.

1.3.3. The Norwegian Function Assessment Scale

The Norwegian Function Assessment Scale (NFAS) is an instrument for self-report that was developed by an expert group in social insurance in 2000 to assess the need for rehabilitation, adjustment of work demands among sick-listed persons as well as the rights to social security benefits (62). The scale comprises 39 items derived directly from the activities/participation component in the ICF (63). The items are relevant for assessing physical and mental functioning in working life, some relating to activities of daily living.

The NFAS starts with the question 'Have you had difficulty doing the following activities during the last week?' and respondents originally self-reported the 39 functional activities using a four-point scale (score range 1-4): no difficulty, some difficulty, much difficulty, could not do it. Later, a version with five all-point defined scale (score range 1-5) has been developed in order to be more congruent with the qualifiers in the activities/ participation component of the ICF: no difficulty, mild difficulty, moderate difficulty, much difficulty and could not do it.

The main application of the NFAS is likely to be social insurance, and the original four-point scale version of the NFAS was tested for construct and convergent/divergent validity against SF-36 and COOP/WONCA, and for utility in a random sample of 386 persons sick-listed for six weeks (62). Based on the results of principal component analysis using this data, the 39 items formed seven domains: Walking/standing (7 items), Holding/picking up things (8 items), Lifting/carrying (6 items), Sitting (3 items), Managing (7 items), Cooperation/communication (6 items), Senses (2 items). The first four and the last three domains are intuitively grouped into physical and mental domains respectively. These domains have evidence for validity in sick-listed persons, and a principal component analysis based on data from a general population would probably yield somewhat different results. Domain scores are calculated by adding the item scores and dividing by the number of items completed. NFAS total scores are calculated by adding all 39 item scores and dividing by the number of items completed. Low scores indicate good functional ability.

1.3.4. The structured Function Assessment Method

Studies have shown that the GPs often have difficulties in adopting a functional approach in relation to sickness certification and assessment of work ability (64;65). In 2003, only 35% of the Norwegian GPs met the request for functional assessments in the sickness certification forms (66). A qualitative study using focus group interviews with GPs showed that the GPs were reluctant and reported difficulties in meeting the request of an explicit communication of patient functional abilities (67). The difficulties could be due to lack of training and guidelines, as well as confusing terminology and insufficient knowledge of specific occupational demands. In addition, the GPs' procedures for functional assessments are usually non-standardised and strongly influenced by their personal and professional interest in functional assessments and working life in general (67).

Based on these findings, a structured method for functional assessments for persons with long-term sick leave in general practice was developed. The purpose was to provide a tailor-made, structured method for GPs in busy and ordinary primary care practices. The method was based on experiences with functional assessments in England, Denmark and the Netherlands, and was designed to be appropriate for assessing and communicating functional ability information along with suggestions for workplace adjustments to local social security officers and employers.

The method consisted of four elements (see Table 1-1). Before the consultation, the sick-listed persons reported their functional abilities on the NFAS and their work ability on a single item. Furthermore, the sick-listed persons reported work exposures and perceived stressors at work by filling in the Work Description Form either at home in advance or at the GP clinic. During the consultation, the GP independently assessed the patient's functional abilities on basis of the two forms, the patient's medical history, clinical findings, and motivation. The assessment was formalised as the Function Assessment Report.

Table 1-1. The four elements of the structured Function Assessment Method

	Name	Description	Appendix
1.	The Norwegian Function Assessment Scale + Work ability	39 physical and mental functional abilities with relation to working life and one work ability item	1
2.	Work Description Form	A simple scheme made by the project group for self-reporting work tasks and perceived exposures	2
3.	Key questions	Six questions put together by the project group concerning the person's resources, own goals and motivation for rehabilitation and return to work	3
4.	The Function Assessment Report	Developed by the project group to constitute a template for functional ability reporting	4

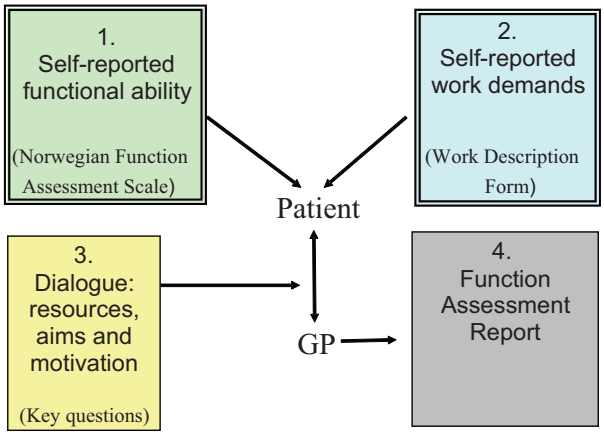
The first page of the Function Assessment Report included putting crosses in boxes for functional abilities/disabilities, the reported work ability level and perceived stressors at work; directly derived from the two self-reported forms. If the patient reported much difficulty or could not do it on one or more items within a NFAS domain, the GP ticked off the box for functional disability for the actual domain. Whereas the GP ticked off the box for functional resources for a domain if the patient reported no difficulty for all items within the actual

domain. When the patient claimed to have functional resources, the GP could ask the patient about how much they thought they could perform within this functional ability domain.

On the second page of the Function Assessment Report, the GP indicated patient resources, possible influences of medical treatment, protective needs (if any) and suggested workplace adjustments. Finally, the report was signed by the GP and the patient before being mailed to the employer and the social security office.

The Key questions represented information that required confidentiality, and information from this dialogue should not be included in the Function Assessment Report or in other ways be transferred to the employer or the social security officer. This dialogue information was meant to be useful for the GP as a basis for discussing realistic rehabilitation goals and how to achieve them. Due to patient confidentiality, no health information or other information not relevant for the functional ability reporting should be included in the report.

Figure 1-2. Model of the structured function assessment method



1.3.5. Brief description on how functional ability/work ability is assessed in England, Denmark and the Netherlands

Work Capability Assessment (WCA) in England

A new procedure for determining entitlement to incapacity benefits was introduced in England in 1995 (54). The Personal Capability Assessment (PCA) (formerly called All Work Test) was developed to provide an objective and impartial assessment of functional limitation. The PCA was considered to be a valid and reliable measure of work ability (55), but few

scientific studies using the PCA have been conducted (19). A revised version applied from October 2008, the Work Capability Assessments (WCA) intend to assess what the individuals can do rather than what they cannot, and identify what personal support they might need in their move towards work (68;69). The WCA should be applied within the first 13 weeks of claiming Employment and Support Allowance, and the assessment is carried out by specially trained nurses and doctors at a Medical Examination Centre (69).

The WCA is made up of three parts. The first part, the assessment of limited capability for work, is a self-administered questionnaire including activities relevant to physical, mental, cognitive and intellectual function assessment. Descriptors within each activity with associated scores are given. The second part is the medical assessment of limited capability for work-related activity. For persons awarded with a score of 15 or more in any physical and/or mental activity or on a combination of activities, will be entitled to Employment and Support Allowance and considered as having limited capability for work. Most of these persons will be placed in a Work-related Activity Group and will take part in work focused interviews and have access to a range of support helping them prepare for suitable work. However, there are also a further eleven activities which are considered to determine if a person has limited capability for work-related activity, who will be placed in a Support Group. The third part is a work-focused health-related assessment, identifying the person's perceptions about work and barriers to work as well as any appropriate health related intervention.

The specially trained nurse or doctor will consider all the information and provide an advice to the benefit decision maker. The WCA will continue to be applied at regular intervals during the life of a claim to ensure the conditions for entitlement are maintained (69).

The assessment of work ability [Arbejdsevnemetoden] in Denmark

As part of welfare reform in 2003, the new 'Work Ability Method' [Arbejdsevnemetoden] was introduced to ensure a systematically description and assessment of work ability carried out by community executive officers when the time limit for receiving sickness benefits is reached (70-72). The main element of the Danish Work Ability Method is the dialogue and interview based resource profile, which consists of 12 elements and represents a tool for both describing, assessing and developing the work ability as well as for identifying the person's sources and barriers in relation to the labour market. During the resource profile process, the executive officer shall be in dialogue with, and receive opinions and suggestions from, various other co-operation partners in order to make a plan of action that is regularly followed

up by the executive officer. Thus, the assessment of work ability is not the responsibility of the GPs. After the assessment, the executive officer and the sick-listed person together seek to find one or more potential jobs that the person can aim for, by matching the resources against vacant posts in the labour market.

The Danish public authorities have been delegated a great responsibility and enhanced incentives to apply active rather than passive instruments, such as vocational rehabilitation and subsidised employment, in order to increase the employment of disabled people. At the same time the responsibility of employers is limited, as their only contribution to the financing of sickness and disability benefits is the first two weeks of sick leave. In addition, the dismissal of sick-listed workers is easy, and the participation of employers in the integration of disabled people is voluntary (73).

Claim Beoordelings- en Borgingssystem in the Netherlands

There exists a medical criterion for benefit allowance in the Netherlands, but at the same time, they have for several years also used functional assessments as a tool for rehabilitation and to calculate disability pension amounts. The Dutch GPs and medical specialists have declined to assess work ability since early 20th century, so a number of positions for insurance doctors, to inspect the legitimacy of sick leave, have been created. The Functie Informatie System was used from 1986 until 2002, then it was replaced by the Claim Beoordelings- en Borgingssystem, which also included psychosocial dimensions (74). A new list of dimensions for the functional assessment was made, Functionele Mogelijkheden Lijst. This list included six dimensions (mental functioning, social functioning, adaptation to physical environment, dynamic motion, static work and working hours) that was further split into sub-dimensions with terms derived from the ICF (7). There are four descriptor degrees of reduced functional ability in relation to what is regarded as normal. The preset normal values represent the level of functional ability of all healthy persons between the age of 16 and 65 years old and examples from activities of daily living are given. Thus, the assessment of the person's abilities is not in relation to previous abilities or the present work demands, but in relation to a predefined level, that is regarded as normal (74). The Functionele Mogelijkheden Lijst is filled in on computer by all sick-listed persons, and works as a checklist for the doctor. The Claim Beoordelings- en Borgingssystem produces a list of jobs that the sick-listed person theoretically can perform with his of her functional ability. A vocational expert uses this list of jobs to identify jobs that the sick-listed person actually can perform and that are available on the labour market (75).

In 2002 new rules were implemented to increase and specify the obligations of employees and employers during sick leave; an action plan and a reintegration report. If they do not fulfil these requirements, the employer's obligation to pay wage during sick leave may be extended and the employee may lose his or her protection against dismissal or the wage payments may be stopped.

The Dutch disability policy is characterised by employers having considerable obligations for the integration of disabled people into the labour market. Employers' dismissal of sick-listed employees is difficult and costly, and individual employers are obliged to finance sick leave for up to two years and disability benefits for up to five years when an employee becomes disabled. Employers have the opportunity to sign contracts with insurance companies to cover these expenses (73).

1.4. Functional assessments and general practitioners

1.4.1. The Norwegian general practitioners' role in assessing functional ability

Norwegian GPs issue about 81% of the initial sickness certifications (76). The GPs are obligated, by § 8-7 in the National Insurance Act (77), to provide an assessment of the employee's functional ability in relation to sickness certification and work disability. The assessment should be done in collaboration with the employee, and sometimes also with the employer. According to § 8-8 in the National Insurance Act (77), it is the duty of the sick-listed person to provide information about his or her own functional ability to the employer and the local social security officer. Thus, a social security officer or an employer may request the GP to provide functional ability information, but so far, there has been a lack of structured instruments available for the Norwegian GPs to carry out this task.

For most Norwegian GPs, assessing a person's function has been an implicit part of their practice in relation to social security claims, whereas at present an explicit communication of functional abilities is required. This new focus on functional ability implies a shift in the GPs' attention from patient symptoms, problems and limitations into resources, possibilities and coping. The assessment often relies on the GPs' judgement, which is based on the patient's verbal case history and findings from clinical examination. As previously mentioned, the assessments are usually non-standardised and affected by the GPs' knowledge and interest (67). Some initiatives have been put into effect in order to standardise the functional

assessment by making it an explicit and useful part of the sickness certification form for The Norwegian Labour and Welfare Service (NAV) and the employer, e.g. the Simplified Functional Assessment [Forenklet funksjonsvurdering] and the new sickness certification form in 2004.

New rules for sickness certifications were implemented by July 1, 2004. This implied that the GPs were requested to use part-time sick leave certifications more often, and to consider part-time sick leave before active sick leave. As a result, Norwegian GP sick-listing practices changed considerably with fewer sick leave cases, shorter duration of sick leave episodes, increased prescription of part-time sick leave, and decreased prescription of active sick leave (78). On the other hand, the percentage of initiated vocational rehabilitations on the national basis remained reasonably stable (79). The new rules also involved another requirement for the GPs: to fill in an expanded medical certificate for sick leave episodes longer than eight weeks, in which no work-related activities are undertaken, in order to certify that there are medical reasons for the non-activity.

During functional assessments the GPs should behave as neutral, medical experts. This is in contrast with the role the GPs normally have as medical doctors in patient consultations, and such role conflict experiences have been described by several authors (80-82).

1.4.2. Changing the general practitioners' behaviour

According to Rogers' theory of diffusion of innovations (83), knowledge represents the first stage in a process of adopting new ideas or changing one's behaviour. The second stage implies a change in one's attitudes towards the innovation before a possible behavioural change can take place. As new knowledge disperses relatively quickly in medical societies, the challenge is getting the GPs to translate their knowledge into action and behavioural changes (84). The complex process of changing one's clinical behaviour demands both energy and active involvement, and proceeding as usual often means to follow the line of least resistance. To start the process of change, sufficient motivating factors must be present, and in a supportive environment, the intentions may be transferred to action. Different forms of rewards or restrictions, like changed economical conditions or social acceptance and belonging, are among the strongest facilitators for a behavioural change (84). Personal motivation along with an understanding of the need for a change are requirements in order to get a behavioural change with lasting effects to take place (85).

Changing or forming an attitude for the innovation represent the second process stage for adopting new ideas or for changing behaviour (83). At this stage, self-efficacy may play an important role, as self-efficacy expectations can affect the initial decision to perform behaviours and the behavioural setting. Further, the expectations can also affect the effort expended and the time the individual will persist in the face of obstacles and aversive experiences (86).

Previous research has shown inconsistent success in changing GP behaviour, but several reviews have concluded that interactive multifaceted interventions, preferably including an opportunity to practice skills and the use of follow-up sessions, are more likely to be effective in changing physician performance compared to passive single interventions (87-89). GP behaviour was changed in some, but not all respects, in a randomised controlled trial (RCT) with structured GP assessments of persons with long-term mental illness (90). In contrast to this, an earlier Norwegian study promoting the prescription of active sick leave for persons with low back pain, showed that the GPs were insignificantly susceptible to the intervention (91). The authors of this study suggested that it might be uncovered barriers at social or cultural level that impeded a GP behavioural change; despite the fact that the GPs were positive in relation to active sick leave. Furthermore, the authors pointed out that there are no strong traditions for cross-professional routine health care services in Norwegian communities. The different players perform their work in a segregated style with a minimum level of communication (91).

2. Study context, rationale and aims

2.1. The study context

This doctoral work has been part of a larger project on functional assessments, 'Functional Assessments - Health Providers' Responsibilities in a More Inclusive Working Life', which was financed by The Ministry of Labour and Social Inclusion for the period 2004-2007 (18;92). The project consisted of five sections whose two primary aims included: 1) critically evaluate and systematise existing knowledge of functional assessment, and 2) develop and test new assessment methods. Furthermore, the project sought to create a firmer foundation and order for function based rehabilitation of sick-listed persons, and a more reliable basis for the assessment of disability pension rights. The long-term aim of the project was to secure the individual's work ability and employment.

In one of these five sections, there was a close collaboration with The Ullensaker Study, which is a population based epidemiological cohort study starting in 1990 with follow-up data collections in 1994 and 2004. The Ullensaker Study is financed by the University of Oslo and Trygve Gythfeldt Fund, and its primary focus is on musculoskeletal pain (12).

2.2. Rationale

There is an increasing interest in instruments that assess health status, functional status and health related quality of life. The ICF-based Norwegian Function Assessment Scale (NFAS) represents an instrument for obtaining self-reported functional ability data with items considered relevant in relation to work. Although the main application of this scale is likely to be social insurance, normative data on functional ability from a population would be of great interest and help for interpreting levels of functional ability. Thus it was chosen to include the NFAS in The Ullensaker Study 2004.

The NFAS had previously been validated in a sample of sick-listed persons, but not in a population based sample, and the test-retest reliability of the scale had not yet been examined. At the same time, it was chosen to develop a five all-point defined scale version of the NFAS, which would be more congruent with the qualifiers in the activities/ participation component of the ICF. The participants were randomised to receive either the original four-point scale or the new five-point scale in order to evaluate which version that should be preferred for future use by comparing the data quality and the validity.

The increased focus on explicit communication of functional abilities in relation to long-term sick leave represents a challenge for the GPs as they lack specific training, guidelines and structured tools to meet with this request. Previous research has shown that this stressed the GPs' knowledge of the patient's work assignments, their ability to map them, and weigh them against the patient's functioning (67). Standardised forms for assessing functioning, quality of life, and health have long traditions in medical science, but most are aimed at assessing intervention effects or for studying population groups. Consequently, they are not suitable for discriminating persons in need for work demand or workplace adjustments or social security benefits. There exists a number of instruments for measuring function, but these instruments are often focusing on physical capacity and measure physical functions like muscle strength, range of movement in joints, or heart and lung capacity. However, physical measure results like these do not easily transfer into an assessment of functional ability. Therefore, a tailor-made, structured method for GP functional assessments in busy and ordinary primary care practices was developed by the project group.

2.3. Aims

The first aim of this study was to obtain population based normative functional ability data, and further validate and test two versions of the NFAS in a population based sample. The second aim was to implement structured GP functional assessments for persons with long-term sick leave in general practice in a cluster RCT.

2.4. Specific objectives

1) To obtain population based normative functional ability data using the NFAS
2) To examine the data quality, internal consistency, validity, and the test-retest reliability of the NFAS in a population based sample
3) To compare the original four-point with the new NFAS five-point scale version by evaluating the data quality, internal consistency, and validity
4) To implement structured functional assessments in general practice and assess effects on GP knowledge, GP attitudes, and GP self-efficacy related to functional assessments as well as GP knowledge of patients' workplaces, work tasks, and perceived stressors
5) To implement structured functional assessments in general practice and assess effects on the GPs' sick-listing practice by analysing the duration of patient sick leave episodes and GP prescription of part-time sick leave, active sick leave, and vocational rehabilitation

3. Materials and methods

3.1. Paper I and II: The Ullensaker Study 2004

Background and design

The Ullensaker Study is a population based epidemiological cohort study with musculoskeletal pain as the primary focus and with two preceding surveys conducted in 1990 and 1994. Ullensaker is a rural community 40 kilometres northeast of Oslo, with 23 700 inhabitants in 2004. In 1990, it represented a typical Norwegian community with respect to demography. Many commute to Oslo for work, but in the last decade, Ullensaker has expanded due to the building of the new Oslo Airport Gardermoen. This has led to some minor changes in demographic factors, as many younger people and people with further education have moved into the community. However, there are still no major differences between the population of Ullensaker and the population of Norway with respect to demographic characteristics (93). The Ullensaker Study 2004 provided an opportunity to collect normative data as well as further validate and test the Norwegian Function Assessment Scale (NFAS).

Participation and randomisation

In the 2004 survey postal questionnaires were sent to all inhabitants in Ullensaker municipality in the birth cohorts 1918-20, 1928-30, 1938-40, 1948-50, 1958-60, 1968-70 and 1978-80 (n=6108). The birth cohort 1978-80 was approached for the first time in 2004. Persons who had moved out of Ullensaker during 1990-2004 as well as new inhabitants in Ullensaker were included in 2004. Information on residence was given by the Population Register. One written reminder was sent after eight weeks.

The sample was computer-randomised by an external company to either the four-point or the five-point scale version of the NFAS. Fifty-five participants in the birth cohort 1968-70 randomised to the four-point version were erroneously mailed the five-point version. Hence, the subsamples differed significantly regarding age ($p < 0.05$), but not on any other background variables. Excluding the birth cohort 1968-1970 did not affect the results.

The response rates for The Ullensaker Study by age and gender are given in Table 3-1. For Paper I, only respondents with the four-point version of NFAS (n=1620) were included, whereas all respondents were included in Paper II (n=3325). However, 15 (1.0%) and 19 (1.1%) respondents had not completed any of the items on the four-point version or the five-

point version, respectively. These respondents were included in the first table in both papers and in Paper 1 for descriptive data regarding education levels, but they could not be included in the statistical analyses using NFAS-variables (Table 3-1).

Table 3-1. Participation rates in The Ullensaker Study 2004 by age and gender. Number of respondents included in statistical analyses in Paper I and II.

		The Ullensaker Study 2004			Included in analyses	
		n	%	Response rate (%)	Paper I n	Paper II n
Females		1824	55	59	899	1806
Males		1501	45	49	706	1485
All		3325		54	1605	3291
Age:	24-26	319	10	36	150	318
	34-36	950	29	52	427	948
	44-46	602	18	54	300	598
	54-56	685	21	65	353	678
	64-66	458	13	69	215	447
	74-76	252	8	64	130	247
	84-86	59	2	36	30	55

The test-retest

For purposes of assessing test-retest reliability, the first 30 that returned the four-point scale version of the NFAS within each of the five youngest birth cohorts were asked to complete the NFAS again after two weeks. The five-point scale version was not included in this test-retest procedure. The two oldest birth cohorts were excluded since they were outside the normal working age in Norway of 16 to 67 years. Individuals reporting no difficulty on all NFAS items were not invited in the retest since possible changes could only be in one direction. As most persons in the youngest cohort reported no difficulty for all questions, there were fewer invited individuals in this cohort (n=17).

Retest questionnaires were sent to 134 individuals, and were returned by 101 (75%). The respondents were significantly older ($p < 0.05$) than the non-respondents, but otherwise comparable.

Questionnaire data

The four-paged Ullensaker Study 2004 questionnaire included among other items the following sociodemographic items: year of birth, gender, education, occupational status and sick leave. Work ability was self-reported using one item, and the NFAS and the COOP/WONCA charts were also included in the questionnaire. Mental distress was covered by the General Health Questionnaire-20 (GHQ-20), and musculoskeletal pain by the Standardised Nordic Questionnaire.

Table 3-2. List of variables used in the statistical analyses in papers I and II

Variables	Paper I	Paper II
Sociodemographic factors		
Gender	X	X
Age	X	X
Education	X	
Disability pension/vocational rehabilitation		X
Sick leave		X
Physical and mental factors		
Functional ability (NFAS)	X	X
Functional ability (COOP/WONCA)		X
Work ability		X
Mental distress		X
Musculoskeletal pain		X

Construction of the variables used in the statistical analyses in paper I and/or paper II

Sociodemographic factors

All sociodemographic factors were self-reported. The question about education included response categories of lower secondary school, upper secondary school (technical), upper secondary school (preparatory), university 1-4 years, university >4 years. Education level was then categorised into three groups: ≤ 9 years, 10 to 12 years and ≥13 years.

Occupational status was reported using the following categories: employed, housekeeping/full-time household work, unemployed, vocational rehabilitation, disability pension, retired or student. For the analyses, this variable was dichotomised into those having a disability pension or rehabilitation benefit due to disease versus all others.

Sick leave was assessed by asking the respondents if they had been sick-listed during the previous year: no, less than 1 week, between 1-8 weeks, more than 8 weeks. This variable was dichotomised into those that reported any sick leave last year versus no sick leave.

Physical and mental factors

The respondents self-reported their functional ability on one of the two versions of NFAS (see section 1.3.3) (62;94;95). Due to limited space in The Ullensaker Study 2004 questionnaire, the domain titles were not included. The mean score for single items, mean domain and mean total scores for the NFAS were calculated for the sample and for each gender separately. For the test-retest analyses, only single items scores for the NFAS were used in analyses.

Functional ability was also self-reported using six COOP/WONCA charts (see section 1.2.3) (21;47;48). In the analyses, only three chart scores were used: physical fitness, feelings and overall health. Each chart has five all-point defined response alternatives with pictorial representations (score range 1-5).

Work ability was assessed by one question 'To what degree is your ability to perform your ordinary work reduced today' with the following response alternatives: hardly reduced at all, not much reduced, moderately reduced, much reduced and very much reduced (score range 0-4) (96).

Mental distress during the last two weeks was self-reported using the GHQ-20 (97), a widely used screening instrument for measuring non-psychotic psychiatric illness in a general population. Items were scored as the original GHQ score in a bi-modal fashion (0-0-1-1). Then the variable was dichotomised into those without mental distress (mean GHQ score <4) and those with mental distress (mean GHQ score ≥4)

Using the Standardised Nordic Questionnaire (98), the respondents were asked to report whether they had experienced any pain or discomfort in ten different body regions during the previous week. This variable was dichotomised into those reporting musculoskeletal pain last week in any body region versus those reporting no pain.

Missing data

For 0.9% of the respondents to the four-point scale version of the NFAS, all 39 items were missing, while 79% had no missing items. The corresponding percentages for the five-point scale version were 1.1% and 82%. All items had more missing values for the four- compared to the five-point scale version. The mean levels of missing values for individual items in the

four- and the five-point version were 3.3% and 2.6% respectively, which represented a statistically significant ($p< 0.01$) difference. It was the same items within both versions that had the highest percentage of missing values. Missing values were omitted in statistical analyses.

Statistical analyses

Paper I:

Internal consistency was assessed by Cronbach's alpha. Test-retest reliability was assessed by calculating total proportions of agreement, weighted kappa (99), and intraclass correlation coefficient (ICC) (two-way mixed model with the measure of absolute agreement). Since the data was categorical, non-parametric tests for independent samples were used to compare subgroups.

Paper II:

Data quality: The two versions of the NFAS were compared for levels of missing data, and floor and ceiling effects, which were expressed as percentages.

Tests of scaling assumptions: Internal consistency was assessed by item-total correlation and Cronbach's alpha. Item-total correlation coefficients should meet 0.40 standard. Cronbach's alpha was considered acceptable for group comparisons when the coefficient exceeded 0.70 (100). Item-discriminant validity was assessed by analysing correlations between the items and their domains (item-total) and between the items and the other domains (item-other) to see if the former was at least two standard errors higher than the latter, thereby indicating definite scaling success (101).

Construct validity: We hypothesised that scores from conceptually related domains of NFAS would correlate higher than scores of unrelated domains. We also hypothesised that NFAS scores would correlate higher with conceptually corresponding aspects of the COOP/WONCA, GHQ and work ability than with non-corresponding aspects. Correlation coefficients among measures of the same attribute should fall in the midrange of 0.40 - 0.80 (25). It was hypothesised that those having a disability pension or rehabilitation benefit due to disease and those reporting being sick-listed in the previous year, would report lower functional ability. We also analysed domain scores for those reporting musculoskeletal pain last week, but no mental distress, and compared them with those reporting mental distress, but no musculoskeletal pain. It was hypothesised that females, older persons and persons with shorter education would report lower functional ability than the males, younger persons and

persons with longer education. Since the data was categorical, non-parametric tests for independent samples were used to compare subgroups.

Table 3-3. Samples and outcome measures in Paper I and II

Paper	Sample	Included in analyses (n)	Outcome measures	Measured as
I	Population based	1605	Functional ability by gender, age and education	Mean scores NFAS four-point scale version
			Missing values	Mean percentages
			Internal consistency	Cronbach's alpha
			Test-retest reliability	Total proportions of agreement, weighted kappa, ICC
II	Population based	3291	Functional ability by gender	Mean scores NFAS both versions
			Missing values	Mean percentages
			Floor- and ceiling effects	Mean percentages
			Internal consistency	Item-total correlation, Cronbach's alpha
			Item-discriminant validity	Item-total vs. item-other correlation
			Construct validity	Correlation coefficients, comparing mean scores

3.2. Paper III and IV: Structured functional assessments in general practice

Background and design

A cluster RCT was conducted from March 1st to October 31st in 2005, to evaluate an implementation of the recently developed structured method for GP functional assessments of long-term sick-listed persons in general practice (see section 1.3.4). Previous research has demonstrated that it is challenging to change the health care providers' behaviour (87-89). The multifaceted intervention was therefore designed to target barriers for functional assessments as identified by a preceding qualitative study, a focus-group study on the introduction of functional assessments in Norwegian primary care (67).

The pilot study

In September 2004 four GPs pilot-tested the structured functional assessments on five long-term sick-listed persons each, and gave feedback to the project group. The GPs also tested the main questionnaire and the evaluation score-sheet. The method, the forms and the logistic were also discussed with two GPs with speciality in physical medicine and rehabilitation, and in a focus group consisting of group members from a Continuous Medical Education group for GPs. As a result, some wording changes were done in the forms and the questionnaires to avoid misconceptions and confusion.

Participation and randomisation

With the assistance of the Section of General Practice, University of Oslo, and of local medical consultants, 360 GPs in the southeastern part of Norway were identified and written invitations were sent in November 2004. Of the 360 GPs, 57 (16%) agreed to participate and were randomly assigned to the intervention or the control group according to a computer generated randomisation list made by an independent researcher. The researchers were not blinded to group allocation.

Six intervention GPs withdrew due to work overload (Figure 3-1), one of them after the baseline measurements and the introduction to the intervention, but before including patients. Two, then three GPs in the control group were lost to the two follow-ups for different reasons. One control group GP moved out of the country and two intervention GPs changed job during

summer 2005, so they did not issue any sickness certification, or report patients off the sick list after this. Additionally, one intervention GP was on maternity leave from February 2006.

The intervention GP that withdrew after the workshop was a male with no speciality, slightly younger than the mean age, and had a smaller list size than the mean. According to the baseline values his score on knowledge of functional assessments was similar to the mean GP score, his attitude and self-efficacy scores were slightly higher, and his knowledge scores in relation to work factors were a little lower than the mean scores. The one control GP that was lost to both follow-ups was a young male with no speciality, reporting a smaller list size, but more daily consultations as compared to the mean for the participating GPs.

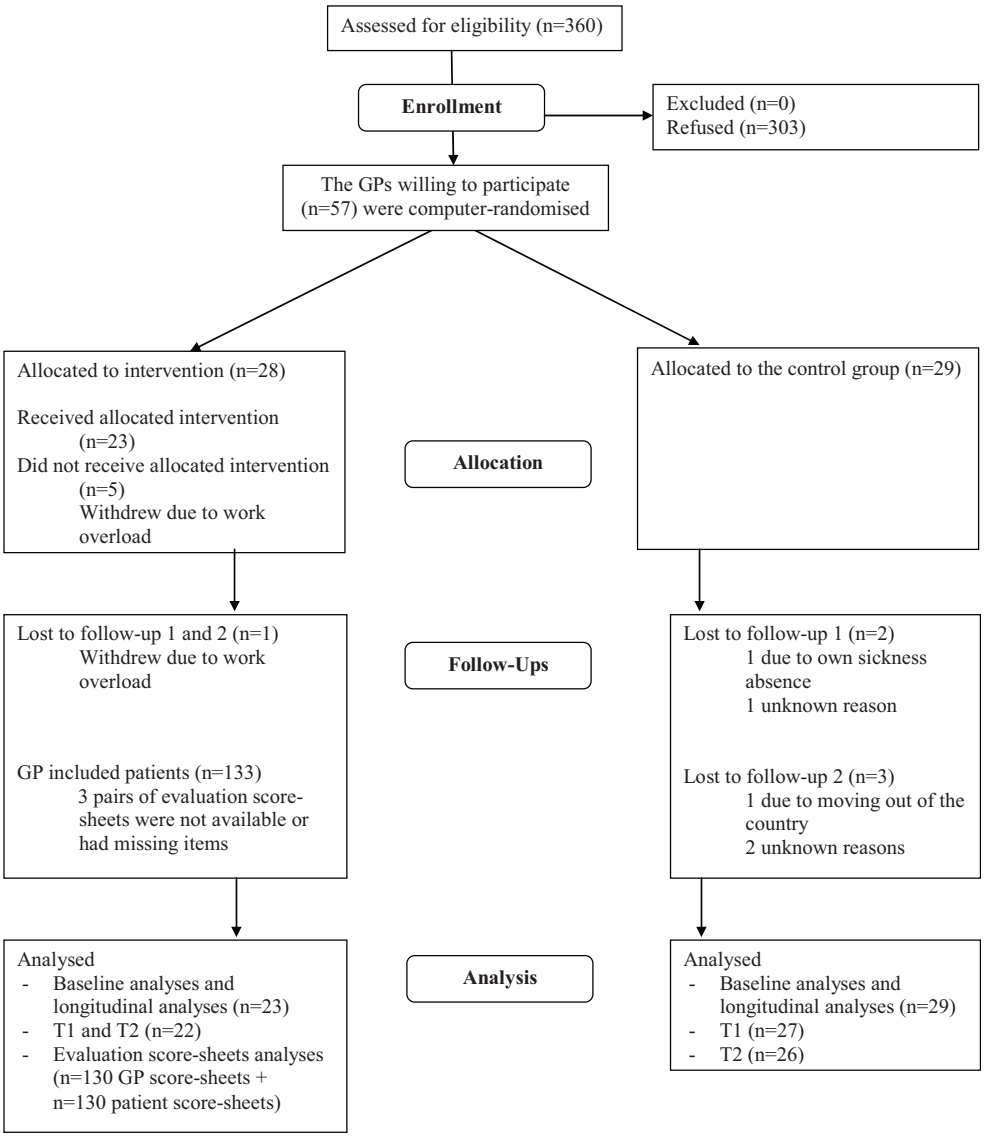
Presentation of the GP sample

No specific information was obtained about non-respondents, but the study sample was compared to all GPs in Norway. The proportion of female GPs (37%) and the GPs' mean age (49 years) in the study sample were slightly higher than the national numbers, but the difference was not significant. The proportion of specialists in family medicine (77%) and the mean list size (n=1285) for the participating GPs were significantly higher than the corresponding national numbers ($p<0.05$) (102).

Close to one in five of the participating GPs worked in solo practices, and a large majority (46%) worked in towns, 30% worked in densely populated area, and only 10% worked in sparsely populated areas. In relation to type of reimbursement, 90% reported per capita reimbursement. The mean number of years since their graduation year was 22 years, and they had on average worked at the present work place for 13 years. The GPs reported that they on average worked for 40 hours per week and had 22 patient consultants per day (range 15 - 35). Seventy five percent of the participating GPs had attended the course: 'The GP role in the more inclusive workplace'.

There was no significant difference between the intervention group and the control group at the start of the study with respect to gender, age, proportion of specialists, weekly working hours, number of daily consultations, or list size.

Figure 3-1. Flow chart of participants through trial



Sample size

Using a table for sample size determination (103) we specified a power of 80% to detect a medium sized difference of 1.2 standardised effect size in relation to knowledge about functional assessments at a significance level of 5%. We found the required sample size to be 22 GPs in each group.

Inclusion and exclusion criteria for application of the intervention

The GPs in the intervention group were requested to apply the intervention on ten consecutive sick-listed persons. The criteria for including a sick-listed persons were: being part-time or full-time sick-listed for between eight and 26 weeks and holding good aspects of a return to work, meaning that the GPs should exclude persons they thought were candidates for permanent disability benefits.

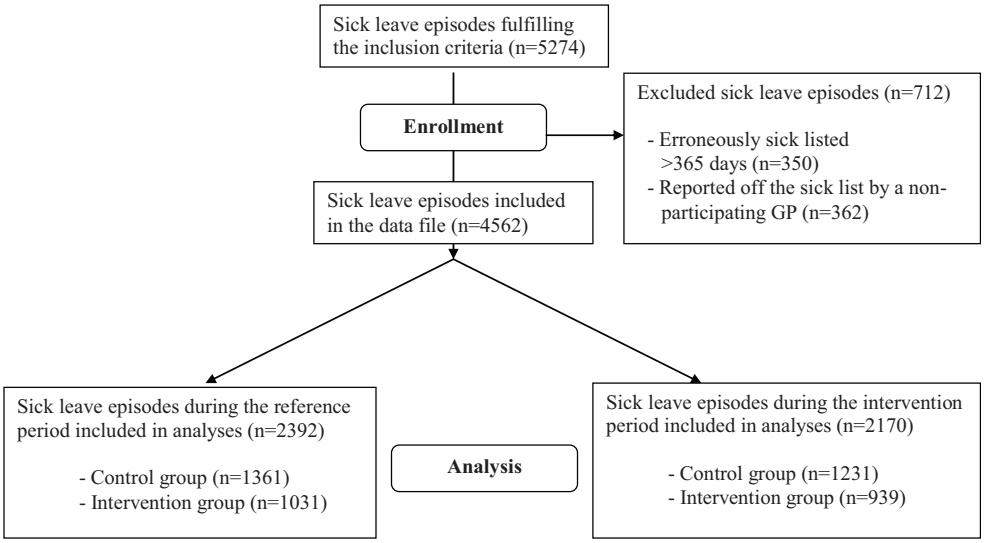
Sick-listed persons included by the GPs

The GPs in the intervention group applied the intervention method on a total of 133 sick-listed persons (two to ten per GP). For these patients, the mean age was 45 years and the percentage of males was 32%, as compared to 42 years and 38% among long-term sick-listed persons on a national basis in the same period (104).

Sick leave episodes from NAV's register

Patient sick leave data was extracted from NAV's sick leave register. For reasons of anonymity, we could not identify the individual patients included in the study. Therefore, all sick leave episodes for the participating GPs were extracted from the register. Of these sick leave episodes, only episodes reaching duration between eight and 26 weeks in the intervention period from March 1st to October 31st in 2005 were included in the analysis file. For historical reference data, sick leave episodes reaching duration between eight and 26 weeks from March 1st to October 31st in 2004 were also included in the analysis file. The sick leave episodes were followed until the person was reported off the sick list or until the limit for receiving sick leave benefits was reached, which is after 365 days. After exclusions (n=712) due to errors in the data file or that non-participating GPs had reported the patient off the sick list (see Figure 3-2), the data file contained 4562 sick leave episodes. For these patients, the mean age was 44 years and the percentage of males was 38%.

Figure 3-2. Flow chart of sick leave episodes through trial



The intervention

The target of the multifaceted intervention was the intervention group GPs, who attended a one-day workshop to learn the structured functional assessment method (see section 1.3.4). The workshop was arranged twice by the project group with about half of the intervention group GPs participating on each occasion. Based on experiences from the preceding study (67), the workshop included a clarification of the terminology and a complex introduction to the structured assessment method. The method was firstly presented orally, and then by a role play conducted by the project group, before the participants practiced their skills in pairs on a case study. The need to practice the structured method as part of the process of the trial was acknowledged, and the project group provided telephone support when needed. The workshop was accredited by the Norwegian Medical Association for continuing medical education points. All intervention GPs received two telephone call reminders during the intervention period, and GPs with low implementation rates received one additional reminder.

The structured functional assessment was expected to take about 40 minutes, and the GPs received reimbursement from The National Insurance Administration (fee L40) for the prolonged consultation during the intervention period. In this clinical trial, the purpose was to test the implementation of the method, and therefore the structured functional assessments were initiated by the GPs and not on request from the employer or the social security officer. The GPs in the control group were requested to assess functional ability as usual.

Data collection

Questionnaire data

The GPs filled in three questionnaires at different time points: the background questionnaire, the main questionnaire and the evaluation score-sheet. The background questionnaire was applied only before the intervention period and included questions about the GPs and their practice (Appendix 5): age, gender, graduation year, speciality in family practice or community medicine, working hours per week, type of clinic, clinic location, type of remuneration, mean number of consultations per day, number on the patient list, and whether they had attended the course: 'The GP role in the more inclusive workplace'.

The main questionnaire was completed by all randomised GPs at three time points (Appendix 6): immediately before the intervention period started (T0), after the intervention period ended (T1), and six months after the intervention period ended (T2). Two written and one oral reminder were given to non-respondents on each occasion. The main questionnaire included 19 items: GP knowledge (item no. 1), GP attitudes (items no. 4 and 7-10) and GP self-efficacy (items no. 11, 13 and 14) related to functional assessments, as well as their knowledge about the workplace, tasks and perceived stressors of their patients (items no. 15-19). The remaining five items were regarded not relevant for this study and were therefore not analysed. All items in this questionnaire were scored along a five all-point defined scale from very poor to very good, or from totally disagree to totally agree.

The third questionnaire was evaluation score-sheets that the intervention GPs and their patients filled in immediately after the consultation (Appendix 7 and 8). They rated the level of the GP's knowledge before and after the consultation using eight items with five all-point defined scales. The patients' evaluations can be seen as validation of the GP's evaluations. Only two of the items (items no. 4 and 8) were included in statistical analyses in paper III, as they more or less covered the other three more specific items.

Register data

Data on sick leave episodes was extracted from NAV's register by a statistician. The file included dates for: the first and the last day of sick leave, the maximum date for receiving sick leave benefits, mean percentage of sick leave degree for the total episode (=100% if part-time sick leave was not used), and prescription of vocational rehabilitation. There was also information about which GP that had issued the patient's first sickness certification and who reported the patient off the sick list. Further, patient information included: gender, age, and ICPC codes for diagnoses. If there were several records of sick leave episodes for the same

individual registered with identical first days of sick leave, these were regarded as one sick leave episode.

Table 3-4. List of variables used in the statistical analyses in papers III and IV

Variables	Paper III	Paper IV
Sociodemographic factors		
GP gender	X	X
GP age	X	X
Patient gender		X
Patient age		X
GP number of daily consultations	X	
Outcome variables		
GP knowledge about functional assessments	X	
GP attitude related to functional assessments	X	
GP self-efficacy for assessing functional abilities	X	
GP knowledge about the workplace (main questionnaire)	X	
GP knowledge about the work tasks (main questionnaire)	X	
GP knowledge about perceived stressors (main questionnaire)	X	
GP and patient knowledge about the workplace and work tasks (evaluation score-sheet) before the consultation	X	
GP and patient knowledge about the workplace and work tasks (evaluation score-sheet) after the consultation	X	
Register variables		
Duration of sick leave episodes		X
GP prescription of part-time sick leave		X
GP prescription of active sick leave		X
GP prescription of vocational rehabilitation		X
Patients classified with a severe disease		X

Construction of the variables used in the statistical analyses in paper III and paper IV

Confirmatory factor analyses using AMOS (105) were used to test the main questionnaire data against hypothesised model structures. As a result, 11 of the included 14 items sum to form three domains (χ^2 (df=59)=71.645, p=0.125): GP attitudes (items no. 4 and 7-10), GP self-efficacy (items no. 11, 13 and 14) and GP knowledge about patient perceived stressors at

work (items no. 17-19). Domain scores for these three domains were calculated by adding the item scores and dividing by the number of items completed. The remaining three items (no. 1, 15 and 16) were used as single items in the following analyses.

The duration of sick leave episodes was defined as the number of calendar days from the first day of the sick leave episode until reported off the sick list. This outcome variable was continuous, with range 57-365 (eight weeks - maximum sick leave). Part-time sick leave was coded as a binary response variable, as whether it was prescribed (mean percentage < 100%) or not (mean percentage = 100%) during the sick leave episode. Active sick leave and vocational rehabilitation were coded as number of calendar days from the first day of sick leave episode to GP prescription of active sick leave or vocational rehabilitation. These were continuous variables with range 57-365. The register included ICPC codes for diagnoses, which were recoded into severe (n=124) versus less severe disease. The severe diseases included malignant neoplasm, cardiac failure, severe head injuries, mental retardation and psychoses.

Missing data

There were few cases of missing values, ranging 0.0 - 1.9% for the main questionnaire and 0.8% for the two evaluation score-sheet items. Two methods of imputing missing values, last-observation-carried-forward (LOCF) and the median imputation, were applied for single items missing and for total questionnaire missing at T1 and T2 (those lost to the follow-ups). This did not change the conclusions, so results from the original non-imputed dataset are presented.

The five intervention GPs that withdrew after the randomisation raises the possibility of post-randomisation selection bias, thus representing a study weakness. As we have no data for these five GPs, we can neither do a drop out analysis, nor follow a true intention–to-treat principle in relation to the GPs.

Statistical analyses

Paper III:

Non-parametric tests for two related samples were used to analyse domain and item score changes in GP attitude, GP self-efficacy and GP knowledge between two time points, whereas the linear mixed model for repeated measurements using Statistical Product and Service Solutions (SPSS) (version 14.0.2) was estimated to assess longitudinal score changes. The linear mixed model was chosen for the two-level longitudinal analyses because it allows

missing items. All variables were treated as fixed effects including the intercept and an interaction variable: time by group. The three domain scores and the three single item scores were used as dependent variables. The covariance (among repeated measures of dependent variables on the same individuals) model was chosen using Akaike's Information Criterion (106), and compound symmetry structure proved best. With compound symmetry covariance it is assumed that the variance is constant across occasions (107), and a close examination of dependent variable correlations between different time points showed low variations. The intervention effect was assessed by the interaction term to analyse if the scores of the two groups differed in time from T0 to T1. To analyse the stability of the intervention effect at the second follow-up, T2 was compared with T1. All estimates in the multivariate models were adjusted for GP gender, age and number of daily consultations.

To assess the potential grading of GP knowledge about their patients' perceived physical, mental and work organisational stressors at the workplace at T0 and T1, Guttman's reproducibility coefficient (108) was calculated. Guttman's reproducibility coefficient shows which fraction of the responses to a set of questions designed to measure one dimension that fits the cumulative pattern. It can be read as the chance to predict correctly the respondent's answer to any given question on the basis of his/her sum score (i.e. the sum of endorsed items in a set of questions). A Guttman's reproducibility coefficient of 1 means that all respondents with a sum score of 1 achieved their one point on the 'easiest' question to agree to, all those who scored 2 points got their points by agreeing to the two 'easiest' questions etc. To conclude that observed data fits a Guttman-scale, the reproducibility coefficient should exceed 0.90 (109).

Evaluation score-sheet data was analysed by two separate linear mixed models for repeated measurements with patients (level 1) nested within the intervention group GPs (level 2). All variables were treated as fixed effects including the intercept, and compound symmetry structure was used as covariance model because before and after scores were correlated. The dependent variables were the GP-evaluated and patient-evaluated knowledge scores after the consultation. Estimates were adjusted for the GP-evaluated and patient-evaluated knowledge scores before the consultation, GP and patient gender and age as well as the number of daily GP consultations.

Paper IV:

Outcome measurement analyses with two-level models were conducted using Stata (version 10.0). Cox proportional hazards survival analysis with standard errors adjusted for GP clusters was used to analyse the duration of patient sick leave episodes and GP prescription of active sick leave and vocational rehabilitation. The patients reported off the sick list before reaching their maximum date and the patients prescribed to active sick leave or vocational rehabilitation, were coded as complete and the others as censored. Part-time sick leave was analysed by a binary response two-level regression model with 4562 sick leave episodes (level 1) nested within the 52 GPs (level 2). All estimates were adjusted for GP and patient gender and age, as well as being classified with a severe disease. The analysis model on sick leave duration included a significant interaction term, 'Group*GP gender'.

Table 3-5. Samples and outcome measures in Paper III and IV

Paper	Sample	Included in analyses (n)	Outcome measure	Measured as
III	GPs in general practice	52 GPs and 130 pairs of evaluation score-sheets	GP knowledge about:	
			- functional assessments	Mean scores
			- the workplace	"
			- the work tasks	"
			- perceived stressors	"
			GP attitude: functional assessments	"
			GP self-efficacy: functional assessments	"
			GP and patient rated knowledge: the workplace and work tasks	"
IV	GPs in general practice	52 GPs and 4562 sick leave episodes	Duration of patient sick leave episodes	Calendar days
			GP prescription of:	
			- part-time sick leave	Yes/no
			- active sick leave	Calendar days
			- vocational rehabilitation	Calendar days

3.3. Ethical aspects

The Regional Committee for Medical Research Ethics and The Norwegian Data Inspectorate approved The Ullensaker Study 2004 and the cluster RCT in general practice. The study was carried out in compliance with the Helsinki Declaration (Ethical Principles for Medical Research Involving Human Subjects Research).

Paper I and II

The collected data was anonymised before any statistical analysis was undertaken. Information about the research project was given to the participants together with the questionnaire. There was no explicit informed written consent, but the return of the questionnaire was presumed as consent to include the respondent in the database and to use the data in different studies.

Paper III and IV

The project's objectives were consistent with Norwegian authorities' intentions for facilitating a quick return to work and promoting work-related activities and part-time sick leave. The project group had regular meetings with the Ministry of Labour and Social Inclusion to review the project status and results as well as for discussing related topics. A reference group consisting of 13 representatives from the Ministry of Labour and Social Inclusion, the labour and the employer unions, NAV, researchers and related professional organisations had two annual meetings with the project group. In this way, we tried to ensure that the project's results would be both useful and used by the decision makers.

In the cluster RCT, the intervention was targeted at the health care providers, rather than directly at sick-listed persons. The intention was to ensure structured GP functional ability reporting and information transfer to the employer and the local social security officer. Project information was mailed to the GPs together with an invitation for participation. All the participating GPs signed a written consent before filling in any questionnaire or attending the workshop.

The sick-listed persons received written information about the project together with the two questionnaires that they filled in before the consultation. The sick-listed persons were informed that all personal data would be treated with confidentiality and that anonymity would be maintained in all statistical analyses and reports. Therefore, no written consent was

collected from the sick-listed persons, but the GPs asked them for a verbal informed consent. The GPs and the sick-listed persons were given the opportunity to withdraw at any time.

The project group attempted to eliminate all possible undesirable or negative consequences for the patient. Despite this, the project had negative consequences for two sick-listed persons as their employers issued dismissals when they received the Function Assessment Report. Although these dismissals later were withdrawn, the incident was undoubtedly a negative experience for the two sick-listed persons. Based on this experience, the project group recommend that the Function Assessment Report should only be sent to the local social security office in future use of the Functional Assessment Method.

4. Main results

4.1. Paper I: Functional ability in a population: normative survey data and reliability for the ICF based Norwegian Function Assessment Scale.

The Norwegian Function Assessment Scale (NFAS) was included in The Ullensaker Study 2004 questionnaire to obtain normative population based data and assess the reliability of the four-point scale version of the NFAS. The questionnaire was mailed to all inhabitants in seven birth cohorts, and 54% (1620 persons) returned the questionnaire. The first 30 respondents in each birth cohort were asked to complete the NFAS again at two weeks, and 75% (101 persons) returned the retest NFAS.

Non-respondents were more likely to be male and young (24-26 years old) or very old (84-86 years old) compared to the respondents ($p<0.001$). Items had low levels of missing values and responses were skewed towards no difficulty. Thirty-three percent of the respondents reported no difficulty for all 39 functional activities. Females, older persons and persons with lower levels of education reported more functional problems than their respective counterparts ($p<0.05$). The age gradient was most evident for three of the physical domains. For females aged 24-56 and males aged 44-76, a clear education gradient was present for three of the physical domains and one mental domain after adjusting for age and gender.

Cronbach's alpha ranged from 0.67 to 0.91 for the domains and was 0.95 for the total score. Test-retest reliability was acceptable with high proportions of absolute agreement, ranging 0.68 - 0.97. Weighted kappa and ICC values ranged from 0.38 (fair agreement) to 0.83 (almost perfect agreement), and 0.79 (substantial) to 0.88 (almost perfect), respectively.

4.2. Paper II: A randomised comparison of a four- and a five-point scale version of the Norwegian Function Assessment Scale.

The Ullensaker Study 2004 sample was randomised to either the original four-point or the new five-point scale version of the NFAS. This allowed for a comparison of the two versions in relation to data quality, internal consistency, and validity.

The response rates for the two versions were similar (54% and 55%). Both versions had low levels of missing values, whereas the four-point version had more missing values for all items than the five-point version. The four-point version had larger floor- and ceiling effects compared to the five-point version. Scaling assumptions (item-total correlation, item-discriminant validity, and Cronbach's alpha) were acceptable for both versions, although the five-point scale version preformed slightly better than the four-point scale version. Construct validity was acceptable for both versions, demonstrated by correlations with instruments assessing similar aspects of health and comparisons with groups of individuals known to differ in their functioning according to existing evidence.

4.3. Paper III: Implementing structured functional assessments in general practice for persons with long-term sick leave: a cluster randomised controlled trial

360 GPs were invited to participate in a cluster RCT, in which a structured method for GP functional assessments was implemented for persons with long-term sick leave. The aim was to assess intervention effects on important GP parameters. The 57 (16%) GPs willing to participate were randomly assigned to an intervention or a control group. Data was collected before, after, and six months after the eight months intervention period. Evaluation score-sheets were filled in by both the intervention GPs and their patients immediately after the consultation to evaluate the GPs' knowledge of patient work factors.

The GPs in the intervention group applied the intervention on a total of 133 sick-listed persons (two to ten per GP). The intervention GPs reported increased knowledge about functional assessments (B: 0.56, 95% CI (0.19, 0.91)), and increased knowledge about their patients' workplace (B: 0.75, 95% CI (0.35, 1.15)) and perceived stressors (B: 0.55, 95% CI (0.23, 0.88)) with sustained effects at the second follow-up. The GPs' self-efficacy for doing functional assessments also increased (B: 0.90, 95% CI (0.53, 1.26)). No intervention effect was seen in relation to GP attitudes. Both before and after the intervention, the GPs were most informed about physical stressors, and less about mental and work organisational stressors (Guttman's reproducibility coefficient: 0.95 and 1.00). After the consultation, both the intervention GPs and their patients reported that the GPs' knowledge about patient work factors had increased (GP B: 0.60 (95% CI: 0.42, 0.78); patient B: 0.50 (95% CI: 0.34, 0.66)).

4.4. Paper IV: Structured functional assessments in general practice increased the use of part-time sick leave: a cluster randomised controlled trial

Of 360 invited GPs, 57 (16%) were willing to participate in a cluster RCT with structured GP functional assessments of persons with long-term sick leave in general practice. The participating GPs were randomly assigned to an intervention or a control group for an intervention period of eight months. Outcome measures included register based data on duration of patient sick leave episodes and GP prescription of part-time sick leave, active sick leave, and vocational rehabilitation.

The GPs in the intervention group applied the intervention on a total of 133 sick-listed persons (two to ten per GP). Sick leave data was extracted from the sick leave register of The Norwegian Labour and Welfare Administration. The number of sick leave episodes included in analyses was 4562. The mean age for these sick-listed persons was 44.3 years and the percentage of males was 38%.

No significant intervention effect was found in relation to the duration of patient sick leave episodes (HR: 0.89, 95% CI (0.79, 1.01). For both groups the mean duration was reduced by five days from the reference period to the intervention period. The intervention GPs prescribed part-time sick leave significantly ($p<0.05$) more often during the intervention period than the control GPs (OR 1.3, 95% CI (1.06, 1.68)). The proportion of part-time sick leave increased significantly ($p<0.001$) for both groups from the reference to the intervention period, 48% to 63% and 48% to 56% for the intervention and the control group, respectively. Significantly ($p<0.05$) less active sick leaves were prescribed by the intervention GPs compared to the control GPs (HR: 0.65, 95% CI (0.43, 0.98)). For both groups, the proportion of active sick leaves was reduced from the reference to the intervention period, from 9% to 5% and from 10% to 7% for the intervention and the control group, respectively. Vocational rehabilitation was initiated in only a small number of sick leave episodes in both groups (3.8%), and there was no intervention effect (HR: 1.04, 95% CI (0.63, 1.70)).

5. Discussion

5.1. Methodological considerations

The following section concerns issues regarding the study design and potential threats to internal and external validity. Advantages and disadvantages for the chosen design will be the first issue followed by internal validity and external validity. The discussion of validity uses the concepts and dispositions of Sackett (110), Kleinbaum, Kupper & Morgenstern (111), and Rothman (112). All types of bias mentioned in this section are in accordance with Sackett's bias definitions (110), unless otherwise stated. Since the study design and validity threats are different for the papers, Paper I and II will be discussed separately from Paper III and IV. The main methodological considerations will be introduced initially, and then a more detailed discussion follows.

5.1.1. Paper I and II

The methodological strengths of Paper I and II are represented by a relatively large study sample, a randomised allocation of respondents to the two Norwegian Function Assessment Scale (NFAS) versions, good data quality, and thorough testing of the NFAS validity against other standards. The self-reporting of functional ability represents a study limitation, and the main threats to internal validity are most likely non-respondent bias and attention bias due to the moderate response rate in the Ullensaker Study 2004. As the NFAS was the main instrument in Paper I and II, it will be evaluated in relation to the criteria for selecting patient-based outcome measures by Fitzpatrick et al. (113) and Deyo's criteria for functional status indices (114). The other instruments used in Paper I and II will be scrutinised in relation to validity and the reliability results from previous research.

Study design

Since the purpose of the present study was to obtain normative functional ability data at population levels and to further validate and test the NFAS, we needed a large, representative population sample. Thus, the Ullensaker Study 2004 provided an opportunity to collect such data, since this study was a population based cross-sectional postal study.

Self-reporting represents an easy and feasible way of collecting normative data, since it requires little administration and allows large samples and geographically scattered participants to be included (115;116). Furthermore, postal questionnaires are inexpensive and

less time consuming for the researcher, and the participants can fill in the answers without interference. However, the depth of responses and the possibility to clarify questions, may be limited in postal questionnaires (116). Personal interviews, direct observation or experimental methods could have been done in order to examine functional ability, but these methods are considerably more time consuming than postal surveys, and are thus not feasible for obtaining data at population levels. Further, while performance based measures tend to assess only a single attribute from the domain of interest, self-report measures are capable of evaluating a number of aspects of function in a single questionnaire (117).

The major criticism of self-reported data is its nature of subjectivity. Self-report measures may be subject to a perceptual or belief mismatch as there may be differences in how patients function and how they believe they function (118). However, when a subjective phenomenon is assessed, subjectivity may also be the strength, since the data reflect personal evaluations of the subject of matter (119). Additionally, observer bias will not be a problem when self-reports are used, but there may be a difference between what patients self-report and what researchers or clinicians would conclude from observations or clinical assessments. On the other hand, self-reported and 'objectively' measured functional ability may seem to complement each other as they capture different aspects of functioning (117;120-122). For example, qualitative aspects of functional activities like dexterity, speed, and compensatory movements cannot be captured by self-reports. In the present study (Paper I and II) we were interested in the functional ability aspects as captured by self-report.

Internal validity

Selection bias

Non-respondent bias may be a concern in population studies if the response rate is low. In The Ullensaker Study 2004 the response rate was 54%, which was slightly lower compared to some (123-125), but higher compared to other population studies using the SF-36 (126-128). A non-response rate of 20-40% is regarded typical in epidemiological studies based on postal or face-to-face questionnaires (129), and the non-response rate in The Ullensaker Study was only slightly higher than this.

A number of studies addressing non-respondent bias in postal surveys have shown that respondents tend to be different from non-respondents in terms of behaviour, as well as demographically. Some have demonstrated that women (129-131), older persons (130) and those in higher social classes (129;132) are more likely to respond to health surveys. It has

also been found that non-respondents are more likely to have low education levels (130;131;133) and higher sick leave (129) or disability pension rates (131).

In The Ullensaker Study 2004 women had higher participation rates than men, and the youngest age group had the lowest response rate for both males and females. Compared with national population data (93), the study sample included fewer persons in the youngest and the oldest cohort. Since these two groups are at the opposite ends of the functional ability continuum, the effects on scores might to some extent have been cancelled out. However, more females than males returned the questionnaire, which might have led to poorer ability levels than if all responded. On the other hand, this effect might have been lessened by the high proportion of persons with education at university level in the sample as compared to the Norwegian population (93).

Information bias

The overall response rate for the NFAS and the low frequency of unanswered items are indications that the questionnaire was acceptable to the population based sample and that there were no major interpretation problems. **Recall bias** was probably not a problem for the NFAS items due to the limited time span, since respondents were asked to relate their functioning according to the previous week. The same was true for almost all instruments used to measure the construct validity of the NFAS, which were all related to the preceding one or two weeks, except for the question about sick leave during the previous year. This sick leave variable was dichotomised into no sick leave versus any sick leave, which might partially have masked the potential recall bias for respondents who remembered being sick-listed, but not the exact duration of the sick leave.

The respondents might have systematically altered their responses in the direction they thought was desired by the survey project group, which is known as **obsequiousness bias**. If this bias was present in the study, it is impossible to tell in which direction this might have affected the results. However, the attention to potential musculoskeletal pain and functional ability in the questionnaire might have affected the respondents to become more aware of pain and disabilities. Such **attention bias** could have increased the reporting of functional disabilities and thereby lowered the population functional ability levels.

Degradation and collapsing of measurement scales tend to obscure differences between groups under comparison, which could result in **scale degradation bias**. In paper I and II, education level categories were reduced from five into three categories before the statistical analyses. In paper II, the items: occupational status, sick leave, mental distress, and

musculoskeletal pain, were dichotomised before the construct validity analyses. This was done to simplify the results as well as the interpretation of the results, but some information might have been lost due to scale degradation.

Although all returned questionnaires were scanned electronically, verified manually, and checked for errors, there may still be errors in the data file, which could have led to **misclassification bias** (134). It is also possible that some respondents mistakenly have put crosses in incorrect boxes in the questionnaire, thereby providing misclassified information. If the incorrect responses were randomly distributed, it would not be a problem, although it might have led to reduced validity and reliability.

Confounding

All observational studies may be afflicted by confounding when the observed association between an exposure and an outcome is actually attributable to a third variable. This was not scrutinised in the present study since the statistical analyses were restricted to comparing group means for different items or item correlation analyses. However, it cannot be ruled out that there may be variables, or sets of variables, that confound the results in Paper I. Some of the significant associations between NFAS domains and education levels disappeared when data was split in gender and age cohorts. This could be an indication of an interaction between age and education levels for males and females or be due to an unknown confounding factor.

Appropriateness of statistical analyses

In Paper I and II only non-parametric tests were used since the NFAS produce categorical data. Therefore, threats in relation to assumptions of statistical tests are not very likely. However, multiple comparisons were done in both Paper I and II, which could have increased the likelihood of falsely concluding that a covariation exist when it does not (Type 1 error). In order to counteract this methodological problem, a data splitting procedure has been recommended (135). Furthermore, there exist different correction methods for multiple testing, e.g. the Bonferroni method (103). By applying a Bonferroni correction, which means dividing the significance level (alpha) by the number of hypotheses, to the analyses of gender differences in NFAS item scores, five of the reported differences in item scores would no longer be significant. Furthermore, the p-values for the domain and total NFAS scores for different age groups (Paper I) would no longer have been significant for the three mental domains for females or for the managing domain for males. For the analyses in Paper II, dividing the alpha by the number of hypotheses would mainly have had consequences for

some of the differences in floor effects between the four- and the five-point response version. However, the Bonferroni method is considered to be highly conservative when the number of comparisons are more than five (103).

External validity

Ullensaker as a proxy for Norway

The demographic characteristics of the population in Ullensaker are similar to those of the Norwegian population, and The Ullensaker Study sample is considered fairly representative of the adult Norwegian population. However, there are more commuters who work in Oslo and more persons employed in the military services in Ullensaker than in Norway in general, which may slightly affect the representativeness of the study sample. In addition, the many younger persons and persons with longer education who have moved into the community area during or after building of the new Oslo Airport Gardermoen might have affected the representativeness of the study sample.

International applicability

The normative data on the NFAS obtained in the present study should be used with caution in international comparisons as there may be sociocultural differences between countries in relation to self-reported functional ability (136). This is probably most evident in relation to some distinct NFAS items, which may be more sensitive to cultural interpretations and differences. This includes the items connected to household work (preparing food, cleaning your house, washing your clothes) and grocery shopping (carrying shopping bags in your hands). The domain scores, and in particular the total scores, are probably less affected by these potential cultural differences.

Instruments

The NFAS and criteria for selecting patient-based outcome measures

A report by Fitzpatrick et al. (113) presents eight criteria for patient-based outcome measure: appropriateness, reliability, validity, responsiveness, precision, interpretability, acceptability, and feasibility. For the present study, the most important criteria include: validity, reliability, precision, acceptability and feasibility. The appropriateness of the NFAS is not relevant in the present study as the objectives were to obtain NFAS normative data, and further validate and test two versions of the instrument. Responsiveness and interpretability will be briefly commented.

The **validity** of the NFAS has previously been found acceptable in terms of construct and convergent/divergent validity against the SF-36 and the COOP/WONCA charts in a random sample of 386 persons sick-listed for six weeks (62). In the present study the validity of the instrument applied to a population based sample was assessed and found acceptable. The NFAS discriminated between groups of individuals hypothesised to differ in functioning and correlated well with instruments assessing similar aspects.

The **reliability** of the NFAS was also examined in the previous study with sick-listed persons, and the internal consistency was considered acceptable (62). In the present study the internal consistency was acceptable for both versions of the NFAS (Paper II), as most Cronbach's alpha values exceeded the 0.70 criterion for group data (100). The test-retest reliability (reproducibility) was assessed and found acceptable for the four-point version of the NFAS (Paper I). It represents a limitation of the present study that the test-retest reliability of the five-point scale version not was tested simultaneously, and it should be assessed in future work. It is possible that the five-point version will have higher test-retest reliability, as there is some evidence that five to seven response categories compared to fewer categories increase the reliability (137). On the other hand, as five categories represent more response choices compared to four categories, the test-retest reliability of the five-point version may also be somewhat lower than for the four-point version. There is generally limited evidence for the choice of different scaling methods (25), and the results may be subject to contextual effects which make it difficult to generalise.

As functional health status is likely to show some day-to-day variation, this will naturally be reflected in the responses, and 100% identical scores cannot be expected in a test-retest. However, it is recommended to check whether the sample has experienced underlying changes in health that would reduce the reliability (113). This can be done by asking a transition question at the second assessment, like for instance 'Is your health or functioning in general better, the same, or worse than at the last assessment?'. It represents a limitation of the present study that the retest questionnaire for the four-point scale version did not include a transition question. However, the results in the present study revealed significant changes in mean scores between the test and the retest for only four items, and for most items mean changes were fairly evenly distributed between improvements and deteriorations.

The **responsiveness** of the NFAS has not yet been assessed as the original objective of the NFAS was to identify needs for rehabilitation or workplace adjustments and assess rights to social security benefits. Thus, if anyone wishes to use the instrument as an outcome measure in a clinical trial, the instrument's ability to detect subtle, but clinically important changes

should be assessed in advance. It should, however, be emphasised that since the NFAS was developed to be a discriminating instrument, items were selected to represent important components of a domain and to be universally applicable to respondents. This is in contrast to instruments developed for use in therapeutic trials, in which the purpose is to detect changes over time for each subject. In this context the item selection would have been focused on components of the domain which are amenable to change over time.

The **precision** of an instrument is influenced by the format of response categories (113). In the present study two different response scale versions of the NFAS was tested and compared. While both versions showed acceptable validity, the new five-point response version demonstrated slightly better data quality, internal consistency, and discriminative validity compared to the original four-point response scale version. The precision of the NFAS may also have been influenced by a variety of selection and information bias similar to those discussed earlier in this section.

The NFAS response categories have an implied rank and are considered ordinal in form. This means that in theory, analyses on interval or ratio levels should not be done, but in practice there may be many times when cautious assumption of interval properties with ordinal data does not seriously mislead (113). According to Streiner and Norman (25), more research is needed about the propriety of using interval level statistics when it is not certain that there is a linear relationship of a measure to the underlying phenomenon.

Fitzpatrick et al. (113) point out that an ideal instrument would have equal precision at every level of, for instance function ability. However, research has demonstrated that some instruments have different precision at different measurement levels. This has not been subject to examination in the present study, but with respect to the skewed item response distribution, it may be hypothesised that the NFAS most likely discriminate better among the persons with poor functional ability than among those with good functional ability. The Rasch analysis is claimed to offer a very useful way of examining the precision of instruments (113). The NFAS was not designed to have the hierarchical (Guttman-like) properties that the Rasch methodology tests, and so far, no such analysis has been done.

An approach to the **interpretability** of NFAS item, domain and total scores has been done in the present study by collecting and presenting population based normative data. Scores in future studies may now be compared to the mean scores and standard deviations of the presented population scores.

The **acceptability** of an instrument has consistently been associated with high response rates (113). Postal surveys tend to have lower response rates than personally administered or

telephone interviews (113), and the overall response rate for the Ullensaker Study 2004 was moderate. However, among those who returned the questionnaire, it was only 1% that had not completed any of the items on the NFAS. Furthermore, there were low levels of missing values for most NFAS items on the returned questionnaires, which may indicate that the respondents found the questionnaire acceptable. Nevertheless, there were four items that had somewhat higher levels of missing values, and these items were related to car driving, working in groups, or guiding others in their activities. As the NFAS was developed for persons in working age, some of the items may be less relevant for students, old age pensioners, or other persons not working. The respondent instruction at the start of the questionnaire tells the respondents to draw a line through the questions that are not relevant for them. Due to limited space, this instruction was not included in The Ullensaker Study 2004 questionnaire. Thus, if respondents considered a functional activity irrelevant, he or she would probably have left this item unanswered. Including a not applicable option is likely to reduce the levels of missing responses for all items, in particular for the four previously mentioned items. To further improve the NFAS instrument, the inclusion of a not applicable response option should be assessed. This is particularly important when the NFAS is included as part of a larger questionnaire, or whenever the respondent instructions are omitted due to limited space.

As the layout, appearance and legibility of a questionnaire are thought to have a strong influence on acceptability (113), it is important to pay attention to these factors during the questionnaire design phase. The layout for the NFAS is more or less fixed, but the appearance may be different from one questionnaire to another depending on for instance the available space and the use of colours. The NFAS appearance in the Ullensaker Study 2004 questionnaire might have been perceived somewhat dull and unappealing as the 39 items were fitted into a black and white single page with no domain titles and with the items following one after another.

The **feasibility** of the NFAS is considered acceptable for staff and researchers with regards to the short format and the simple scoring system. NFAS analyses may be based on single item scores and/or domain and total scores. However, the item responses constitute categorical data that are skewed towards the first response alternative, no difficulty. This was not a problem in Paper I and II since non-parametric statistical models could be applied to compare mean scores of subgroups and for correlation analyses. For more advanced statistical analysis models that hold assumptions of normally distributed residuals, e.g. regression analysis, the skewed response distribution may imply a challenge whenever NFAS variables

are used as dependent variables. A log transformation of the NFAS variables might be helpful in this context or alternatively non-parametric regression methods can be used.

The NFAS related to Deyo's five criteria for functional status indices

The self-administered format and the brevity of the NFAS fulfil the practicality criterion for functional status indices (114), and with regards to being comprehensive, the NFAS include both physical as well as mental aspects of work-related functioning. However, if the NFAS had been developed to be more comprehensive, it would be at the sacrifice of the brevity. The NFAS can be completed in ten minutes, which is acceptable for respondents. As commented above, the reproducibility (reliability) and validity of the NFAS have been found acceptable, whereas the responsiveness has not yet been assessed.

The NFAS as compared with the Work Ability Index and the SF-36

The Work Ability Index (WAI) was developed by the Finnish Institute of Occupational Health in the early 1980s for practical use in the occupational health care as an aid to help maintain work ability (56). The SF-36 has previously been briefly described (see section 1.2.3). The NFAS, WAI and SF-36 are all self-administered and assess the person's own concept of ability in relation to both mental and physical aspects. The three instruments are all relatively short, but while the WAI is four-paged and the SF-36 is three-paged, the NFAS can easily fit on a two-paged paper or even on a single-paged paper if respondent instructions and domain titles are omitted.

The WAI contains ten questions, and five of these are related to work ability (current work ability, work ability in relation to physical and mental job demands, work impairment due to diseases, and prognosis of work ability) and three are related to mental resources. The SF-36 consists of 36 questions and is more comprehensive compared to the WAI and the NFAS in relation to aspects of physical and mental health. All 39 NFAS items are questions about specific physical or mental functional activities, whereas only ten of the questions in the SF-36 are related to specific activities. All three instruments include questions about abilities related to leisure time. The NFAS is consistent in relation to response categories for all questions, whereas the SF-36 uses two to six and the WAI three to ten response alternatives.

The other instruments in the present study

To examine the construct validity of the NFAS (Paper II), the following instruments from The Ullensaker Study 2004 were used: the COOP/WONCA charts, the Standardised Nordic

Questionnaire, and the GHQ-20. Additionally, three single items about work ability, occupational status, and sick leave were used.

The Norwegian version of the COOP/WONCA charts was used in the present study to obtain self-reported functional ability in relation to physical fitness, feelings and overall health. Previous studies have demonstrated acceptable validity and reliability (46;49-52), but the validation was done primarily in clinical settings and in primary health care, and is not necessarily generalisable to a population survey.

Information about musculoskeletal pain was obtained by the Standardised Nordic Questionnaire, which has shown acceptable reliability and validity (98). Acceptable levels of correlation between the questionnaire and examination by physiotherapist was shown (138). The testing was, however, restricted to low back pain and done in an occupational setting and is not necessarily generalisable to a population survey.

The GHQ-20 was used in The Ullensaker study 2004 to measure mental distress. It is a well validated and widely used questionnaire (139), and it was translated into Norwegian in 1978 by Tom Andersen, University of Tromsø (140). Acceptable validity and reliability for this Norwegian GHQ-20 version have been demonstrated (141).

The work ability item in the present study was originally developed from a Gradual Reduced Work Ability Scale (142). A previous study using this item has demonstrated high agreement on work ability assessments by the patient and their GPs at the start of new sick leave episodes (96). Further, two studies have shown that the patients' self-assessed work ability predicted the duration in prolonged episodes of sick leave (143) and return to work for persons with back disorders (144). The questions on sick leave and occupational status were formulated by the Ullensaker project group.

5.1.2. Paper III and IV

The main methodological strengths of Paper III and IV are the cluster RCT design, the second follow-up on stability of results, and the use of register based outcome measures. Study limitations are represented by the self-selection of participants and the post-randomisation withdrawal. Furthermore, the lack of psychometrics testing of the questionnaires and the self-reporting of some outcome measures are important methodological considerations that will be discussed in this section along with other considerations and potential threats to the internal and external validity.

Study design

To evaluate the implementation of GP structured functional assessments in general practice, the RCT design was chosen because it is the method of choice for studying the effects of an intervention in clinical research (145). Findings from RCTs are considered to hold a higher degree of evidence than findings from observational studies due to the randomisation process, which intends to ensure that the characteristics are equally distributed between those who receive the intervention and those who do not. The randomisation ensures that any outcome effect is caused by the introduced exposure.

We used only quantitative methods in the present study, but we could for instance have interviewed or observed the GPs and their patients during the consultations. This would have provided an opportunity to explore the feasibility and the usefulness of the intervention method and thereby capture qualitative aspects of the implementation. Due to the limited time for this doctoral work, qualitative methods were not included, but this may represent a potential future study. However, simple post-evaluation telephone interviews with the intervention GPs were done, and the results have been reported elsewhere (92).

The two main reasons for including patient self-report instruments in the structured functional assessment method were: firstly, to involve the sick-listed person in the assessment of his or her own functional ability; and secondly, to be time-efficient for the GPs. The choice of patient self-report is in line with an international trend, in which the importance of the patient's perspective is increasingly recognised by clinicians and researchers (146). Hence, there has been a shift from 'objective' (e.g. range of motion or muscle strength) towards subjective evaluation tools focusing on patients' self-report of physical function, e.g. in low back pain research (116).

There is an ongoing debate about the validity of self-reported questionnaires versus performance based measures to assess physical functioning (147), but some researchers have recommended the use of self-report instruments in relation to functional assessments (114;136). This is due to the low or moderate association between self-reported and performance based function (120;136), and to the weak correlation between physical measurements of muscle strength or range of motion and actual patient behaviour or symptoms (114;120). Previously, performance based measures of functional status have been thought to hold several advantages over self-report measures for both clinical and research purposes, including greater patient acceptability, interpretability, reproducibility, sensitivity to change, and the focus on actual ability rather than presumed capability. However, in a previous study, the performance based measures were not found to be psychometrically

superior related to the questionnaire (148). Further, it has been claimed that measures of flexibility, strength, or timed activities often may reflect non-physical, highly subjective states in addition to the actual physical capabilities (114).

In relation to social insurance, the recent Swedish official report (149) repeatedly point out the importance of the individual's perspective and involvement during description and assessment of the individual's own work ability or functional ability. By allowing involvement, the process will presumably be experienced as legitimate, irrespective of the actual result of the assessment. This is in accordance with our reasoning for using a self-reported instrument in the structured assessment method.

The demand for a system within social insurance to assess functional ability and disability is partly conditioned by the public authorities' need to deal with the so-called moral hazard risk. How can those who have disability due to sickness be distinguished from those that exaggerate their disability in order to receive social security benefits? Although the motivation of the individual should not be an important aspect in relation to work ability assessments, it is important not to forget that self-confidence and motivation represent crucial, if not the most crucial, aspects in relation to whether the person returns to work or not (149).

Internal validity

For RCTs the design minimises threats to internal validity, but there are still some threats that need to be taken into account (150).

Selection bias

Only 57 (16%) of the invited GPs were willing to participate in the cluster RCT. A larger sample is recommended (110), but is difficult to achieve when recruiting GPs (151). No information was obtained about non-respondents, but the study sample was compared to all GPs in Norway (see section 3.2). The participating GPs probably belonged to a group that was more interested in functional assessments than other GPs were. Such self-selection, or **volunteer bias**, was unavoidable and most likely reflects that mainly interested GPs voluntarily seek special skills and utilise structured methods in this domain. This limits the generalisability of the results.

A further threat to internal validity in the present study is the post-randomisation **withdrawal bias**, as five intervention GPs withdrew before the intervention period, which lead to differential attrition rate for the two groups. If they represented the participants who were least motivated, this would result in a systematic selection of only highly motivated GPs

in the intervention group and thereby violate the randomised design. Considering the actual results, this could mean that the differences in GP knowledge and GP self-efficacy levels were overestimated. Since we have no baseline date for the five GPs that pulled out of the study before baseline measurements, no drop-out analysis, nor true intention-to-treat analyses, could be done. The one intervention group GP that withdrew after the workshop, had slightly more positive attitude towards functional assessments and higher self-efficacy baseline scores compared to the mean baseline scores for the study sample.

As the patients' age and gender were known, we could compare our patient sample with the mean age and the gender proportion of long-term sick-listed persons on a national basis in the same period. However, we do not know if any other patient characteristics differed, as the patient sample population in certain practices might be different from the patient population seen in other practices. The reputation of a clinician or practice may draw individuals with specific disorders to them and thereby skew their patient population, a bias known as **centripetal bias**. Since we do not know much about the patients in the present study, we cannot predict how this potential bias may have affected our results.

The patient sick leave was used as an indirect measure of the GPs' sick-listing practice, which was accounted for by the application of two-level data models in the statistical analyses. Since we were not allowed to identify the 133 patients that were assessed by the intervention GPs, we included all sick leave episodes for the participating GPs in the analyses. This might have weakened the intervention results, but in our view it may represent a more valid picture of how the GPs' sick-listing practice was changed.

Information bias

The overall response rate for the main questionnaire and the evaluation score-sheet in addition to the low missing levels may indicate that the GPs found the instruments acceptable and that there were no major interpretation problems (Paper III). An additional indication of acceptance is the absence of written comments in the questionnaires. Further, **recall bias** was probably not a problem in the present study as the main questionnaire concerned factors at present and the evaluation score-sheets were filled in immediately after the consultation.

The outcome measures might not have been capable of detecting clinically significant differences, which is known as **insensitive measure bias**. More sensitive measures may have revealed other important differences between the two GP groups.

As GPs like to be clever (152), the data may be subject to **obsequiousness bias.** If the study sample GPs systematically have responded to the questionnaires in the direction they

perceived or thought was desired by the project group, it may have resulted in an overestimated intervention effect on GP knowledge and GP self-efficacy (Paper III). That no similar effect was seen in relation to GP attitude may either indicate that obsequiousness bias not was a large problem in the present study or it may be due to the already high attitude level found at baseline giving limited scope for further improvement. Still, we cannot completely disregard that obsequiousness bias may have affected the results in Paper III. Furthermore, the GPs might have systematically altered their behaviour in relation to sickness certification practice because they expected that their practice would be scrutinised (Paper IV). However, as the intervention period lasted for several months and the trends in the study were comparable to national trends for sick-listing practice during the intervention period, this is not likely to be a large bias in the present study.

It can not be ruled out that the extra attention the intervention GPs received through the workshop and telephone contacts, affected their questionnaire responses and clinical behaviour compared to the control group. If such **attention bias** was present, it would probably have led to overestimated effects of the intervention.

In experiments requiring participant adherence to an intervention, issues of efficacy become confounded with those of non-adherence, which is known as **compliance bias**. The low intervention rate among the GPs could either indicate that there were few patients fulfilling the criteria, or that the GPs had low compliance in relation to the intervention method. A handful GPs admitted that for some sick leave episodes, despite telephone reminders, they had forgotten to implement the intervention method. Simultaneously, several of the intervention GPs stated that the number of eligible patients, considering the inclusion and exclusion criteria, was lower than expected, because the patients were either reported off the sick list before eight weeks or they were judged having poor prospects of a return to work. A potential low compliance and the low implementation rate might have weakened the intervention effects.

To our knowledge, **contamination bias** was not likely to be a problem in the present study, as none of the control GPs worked in the same clinic or practice as any of the intervention GPs. However, it is not known if any of the participating GPs share a circle of acquaintances and in this manner exchanged knowledge or experiences, which could diminish the differences in outcomes of the two groups.

Confounding

In an RCT, the randomisation procedure is intended to balance out potential confounding factors between the compared groups. However, post-randomisation baseline levels of various characteristics may differ between the randomised groups due to chance, particularly if the sample size was low. In the present study, no significant difference was found between the two groups with respect to the obtained background information. Due to the low sample size it was chosen to include a few variables based on empiric knowledge, e.g. GP and patient age and gender, in all multivariate models to adjust for possible confounding effects. In the analyses in Paper III, the crude and the adjusted estimates were similar indicating that the included independent variables had no confounding effect on the outcome, whereas in the analyses in Paper IV, the estimates changed to some degree when GP and patient gender and age variables were included in the model. It is, however, possible that there might have been residual confounding that was beyond control or simply unknown.

Statistical analyses

Potential threats to the validity of statistical conclusions may include violated assumptions of statistical tests, low statistical power, reliability of our measures and reliability of the intervention implementation (150). The **assumptions** for using linear mixed models with repeated measurements (Paper III) were found acceptable in relation to the distribution of residuals. The binary response regression model (Paper IV) was checked with regards to residuals and outliers and was found satisfactory. To examine the proportional hazards assumption for the Cox proportional hazards survival analyses (Paper IV), different plots (e.g. Kaplan Meyer curves, Nelson-Aalen cumulative hazard estimates) were checked and found acceptable. In the sick leave data file (Paper IV), a small proportion of individuals were registered with two distinct long-term sick leave episodes. It was chosen not to introduce this third level in statistical analyses, but to treat these episodes as if they were independent sick leave episodes. Hence, the assumption of independent observations may have been slightly violated in these analyses.

The number of intervention GPs in statistical analyses after the post-randomisation withdrawal was equal to the minimum required sample size from power calculations in relation to GP knowledge about functional assessments. Thus, **the statistical power** could have been too low to detect significant differences in relation to duration of sick leave episodes between the two groups (Type II error).

The randomisation procedure does not rule out atypical group behaviour as a threat to internal validity (150). **Atypical control group behaviour**, like compensatory rivalry or treatment imitation, makes true differences difficult to detect. Because none of the GPs worked in the same clinic, atypical control group behaviour was not regarded as a problem in the present study. Although the duration of the sick leave episodes for the control group patients decreased and the control group GPs increased their use of part-time sick leave, this was probably due to the new sickness certification rules implemented by the July 1, 2004, rather than being compensatory behaviour or treatment imitation.

Reliability of the intervention implementation relates to whether the intervention was different from one GP to another, or from one patient consultation to the next patient consultation for the same GP. The intervention method was designed to be as structured and standardised as possible, but it cannot be ruled out that there could have been minor differences in how the GPs implemented the intervention. For instance, the interpretation of the inclusion criterion, 'having good prospects of a return to work', could have been interpreted differently by the GPs. Such subjective judgements might have led to skewed method implementation, which may in turn have given the GPs a distorted impression of the method's usefulness related to analyses in Paper III. It is, however, not known in which way this could have affected the results because we do not know whether the GPs may have been restricted or liberal in relation to judging the probability of a return to work. In Paper IV, all episodes of sick leave at the predefined time and with the predefined duration were included in the analyses, regardless of whether the method was implemented or not. Hence, skewed patient selection was not likely to be a problem in Paper IV.

External validity

Representativeness of the participating GPs and their patients

The study sample is representative for Norwegian GPs with regards to age and gender. However, this study sample proportion of specialists and their list size were higher than national numbers. Despite representing a selected group that is more interested in functional assessments than many other GPs are, all national trends with regards to GP sick-listing practices were reflected in our study results. This could be an indication of the study sample representativeness. Further, the sick-listed persons were comparable with long-term sick-listed persons on a national basis in relation to age and gender, but it was not possible to examine their representativeness for any other characteristic. This implies that generalisation

of the study results should be done with caution. Further research is needed to examine if the intervention effects continue and if the study results could be replicated in a larger study.

International applicability

The social security system in different countries is unique, and the international representativeness of the data in relation to effects on patient sick leave (Paper IV) cannot be claimed without caveats. Although GP knowledge, GP attitude, and GP self-efficacy (Paper III) probably are less affected by system differences, they may possibly be affected by sociocultural differences.

Instruments

The GP background questionnaire (Paper III and IV) was a slightly modified version of a corresponding background questionnaire from a previous Norwegian study in general practice (152). The main questionnaire and the evaluation score-sheet (Paper III) were made by the project group since none of the existing instruments were regarded to be appropriate for our purpose. All three questionnaires were critically evaluated by two medical doctors and tested in the pilot study in advance, which may to some extent account for face validity. It should, however, be questioned whether what we measured as GP knowledge, GP attitude, and GP self-efficacy reflected the underlying constructs. It represents a weakness of the study that no psychometric statistical analysis was done for the main questionnaire. However, the high response rates may indicate high acceptability.

We did not conduct a test-retest comparison regarding the main questionnaire and the evaluation score-sheet (Paper III), which represents a limitation of the reliability of these measures. Errors during the transmission of data recording from paper questionnaires to electronic data files could have occurred (Paper III). Efforts to reduce this potential source of error include thorough quality assurance during the transmission process and searches for extreme or incorrect item values in the data file. Additionally, a systematic control of 10% of the data file was done, and no error was detected. The corresponding evaluation score-sheets from the GPs and their patients were successfully matched using the recorded consultation date, patient gender and age, and the GP code that was written by hand by the project group on all questionnaires.

For unknown reasons there were errors in the sick leave data file from NAV since some sick leave episodes (n=350) extended the maximum duration for sick leave in Norway (365 days). These were excluded from the statistical analyses in Paper IV. The other sick leave

variables in the data file were also checked for values outside the defined range, and no error was found.

5.2. Discussion of main findings

This general discussion aims to integrate and further discuss the findings from the four papers included in this thesis. In the present study, the NFAS has been applied as a measure of population functional ability and as a measure of sick-listed persons' functional ability as a part of a structured method for assessing functional ability in general practice. Advantages and disadvantages for these two applications as well as implications for future work will be addressed. Furthermore, experiences related to the implementation of the structured functional assessment method in general practice will also be discussed in the following section. Some of the results regarding the validity, reliability and testing of the two NFAS versions have already been discussed in the previous section. The following three sections are related to the first three specific study objectives, whereas the fourth section relates to the two last specific objectives. In this last section the results are combined into positive and adverse intervention effects along with considerations for future use of the structured functional assessment method.

5.2.1. Application of the NFAS to obtain population based normative functional ability data

For surveys aiming to collect normative functional ability data, it is important that the sample is representative of the general population or the target population. Although the population in Ullensaker may be representative of the Norwegian population to some degree, the data presented in Paper I and II are not national population norms, but they may serve as a basis for the development of such norms or be used as normative population data.

Normative data for the NFAS may be useful in epidemiological research for comparing scores from different samples, for instance with patient sample scores. An individual's particular NFAS domain or total score does not mean anything unless compared with former scores for the individual, with normative data, or with other reference group mean scores and standard deviations. For the five-point scale version of the NFAS, all Cronbach's alpha coefficients exceeded, or were very close to exceeding, the 0.90 standard for individual

comparisons (100), except for the sitting and senses domains (Paper II). Hence, comparisons with individual scores may be done with caution.

Thirty percent of the respondents in The Ullensaker Study 2004 reported no functional disabilities, whereas the remaining respondents reported various degrees of functional disabilities in one or several functional activities. The results showed that self-reported functional ability levels differed according to gender and age. Functional ability levels were also different in relation to length of education, but the relation between age, gender, and education was difficult to disentangle. Thus, we suggest that presentations of normative functional ability data for the NFAS must take age, gender, and education into account.

The NFAS has been translated into English by a native English researcher, and the instrument has also been translated into Icelandic for application in an intervention study with a sample of persons on vocational rehabilitation. Furthermore, there are plans for a German translation of the five-point scale version of the NFAS for use in a population based cross-sectional study about functional health and work ability. The study will be representative of the German population between 20 and 75 years of age, and the NFAS will be included together with other instruments like the WAI and Copenhagen Psychosocial Questionnaire (COPSOQ). This will allow for cross-national comparisons of normative population data.

In future work it may be interesting to collect data from different patient samples. Then the data could be compared with normative population data to describe the disability degree and potential resources. Furthermore, this data could be used to make a description of which NFAS items or domains that characterise the functional ability and disability related to different patient samples. In social insurance research it would be interesting to find out how much the functional ability is reduced for different population groups receiving various social security benefits, e.g. persons on vocational rehabilitation or persons receiving disability pension.

5.2.2. Validity and reliability of the NFAS in a population based sample

Although originally developed for use in social insurance, the NFAS demonstrated acceptable validity and reliability as a population survey measure in the cross-sectional Ullensaker Study 2004 (see section 5.1.1). This finding advocates an application of the NFAS in epidemiological research on work-related functional ability.

The NFAS items represent selected work-related categories from the activities/ participation component in the ICF. However, the NFAS grouping of items is different from

the original grouping in the ICF component. The item grouping in the NFAS reflect results from a principal component analysis from a previous study with sick-listed persons (62). It is possible that a principal component analysis based on data from a general population would yield somewhat different item grouping, but similar analyses for the SF-36 have replicated the hypothesised eight scales in patient and general populations (153;154). Grounded in the belief that the main application of the NFAS is likely to be social insurance, it was decided to keep the domains from the study with sick-listed persons. Therefore, a principal component analysis based on the Ullensaker data was not done.

5.2.3. The four-point and the five-point scale version of the NFAS

While both versions showed acceptable validity, the five-point response version demonstrated slightly better data quality, internal consistency, and discriminative validity. As the responsiveness of the NFAS has not yet been examined, we can only hypothesise that the five-point response version has greater potential to be more responsive to changes compared to the four-point response version because it provides an additional response option. A further argument for the five-point response version, is the fact that the NFAS is derived form the ICF, which uses five-point scales both in the classification (7) and in the core sets (63). For these reasons, the five-point response version is the recommended version for future applications of the NFAS in social security and epidemiological research. Normative data for the five-point version is published in a report written in Norwegian (92) and is available in English at the project's homepage: http://www.med.uio.no/iasam/sosmed/funksjon/. Item and domain mean scores are also presented in Paper II.

It represents an interesting finding that the respondents not seemed to be concerned about the actual wording of the response categories. Since the respondents were randomly assigned to the two NFAS versions, it could be expected that the proportions of respondents choosing one of the three identical response categories would be fairly similar. This was contradicted by the significant differential distribution of item responses for the two versions. Thus, it may be hypothesised that the respondents looked at the response alternatives as a scale with a given number of categories, and that they evaluated their functional ability with little attention to the category wording, but instead on the basis of the number of boxes. Given the variation in the choice of response scales and limited supporting evidence (24), the results produced by this RCT may inform other researchers in their development of future questionnaires.

5.2.4. Structured functional assessments in general practice and intervention effects on important GP parameters and patient sick leave

Structured functional assessment methods have been introduced and are used in other countries (55-57), but to our knowledge, this work represents the first attempt to implement a structured method for functional assessments of long-term sick-listed persons in general practice. The application of an RCT design strengthens the results of the intervention, and the historical reference data on the GP's sick listing practice provides an opportunity to control for other effects, like for instance legislation changes.

Positive effects of the method implementation

The results in the present study were promising in relation to the modest intervention that was implemented: a one-day workshop with some telephone support. The findings indicate that this was enough to change the GPs' clinical behaviour to some extent. Furthermore, it significantly improved GP knowledge of functional assessments and patient work factors in addition to GP self-efficacy related to functional assessments with sustained effects at the follow-up.

Sick leave may theoretically have positive or negative consequences for the individual, but the consequences of sick leave is inadequately investigated (155). Nevertheless, there is relatively broad consensus that sick-listed persons should, when possible, remain in work or return to work as soon as possible. This is because the beneficial effects of work outweigh the risks of work and are greater than the harmful effects of long-term sick leave (156). However, return to work may grow difficult as the length of the sick leave spell increases (155;157), and the tendency of becoming isolated, inactive, and trapped in a negative sick role increases as time passes (158). On the other hand, a sustained or re-established contact with the workplace, employer, and colleagues may prevent the sick-listed persons from getting increasingly distanced from the work sphere of society (159). Hence, part-time sick leave might prevent the negative effects and consequences that are associated with long-term sick leave. Furthermore, the society's economical burden of long-term sick leave may be reduced by an increased use of part-time sick leave. Additionally, as two-thirds of those on full-time sick leave believed that part-time sick leave would have been positive for them (160), there should be scope for a large increase in the number of persons returning part-time to work.

Most GPs expressed satisfaction with the structured tool for performing functional assessments in their clinical work. By adopting this method, the GPs could assess work-

related functional abilities and transfer functional ability information along with suggestions for workplace adjustments to the local social security officer in a standardised and structured manner. This structured information may in turn guide and aid the local social security officer in their work with sick-listed persons' rehabilitation and return to work.

The individuals' own experience of functional ability represent an important aspect in relation to rehabilitation or reintegration to work (17). Furthermore, the direct involvement of the sick-listed person in the assessment of his or her functional ability is regarded as important (149), which the GPs do when they implement the structured method on their patients and ask them to fill in the NFAS and the Work Description Form.

Missing or adverse effects of the method implementation

The results from this study showed that the method implementation did not affect the GP attitude in relation to functional assessments, the duration of patient sick leave episodes, or GP prescription of vocational rehabilitation. Possible explanations are the high GP attitude levels at the project start giving limited scope for further improvement, and that the GP sample size was too small to detect effects on duration of sick leave or prescription of vocational rehabilitation. This is further discussed in Paper III and IV.

The method's end-product, the Function Assessment Report, was intended to be helpful in the reintegration of the sick-listed person into working life, hence the focus was meant to be on the functional abilities, resources, and the persons' potential for work. However, if applied in a different context, the intention of the assessment might be to evaluate the criteria fulfilment for social security benefits. In this context it could be appropriate focus on disabilities. This distinction is in line with a recent Norwegian Official Report about disability pension, which categorised functional assessments as holding either a counselling or a management purpose (15). If, however, these two intentions and uses are mixed, adverse effects may be the result. This was experienced in the present study as two employers issued dismissals for the sick-listed person when they received the Function Assessment Report (see section 3.3). This is in contrary to the intended use of the NFAS in the present study, which was to aid the employers in their effort to reintegrate the sick-listed persons. Therefore, the project group now recommend for future use that the Function Assessment Report only should be sent to the local social security officer to avoid this kind of misuse of the report and the functional ability information. It will then be the sick-listed person and/or the local social security officer's responsibility to inform the employer in relation to the need for guidance and adjustments of the work tasks and the workplace.

The project group have received copies of the Functional Assessment Reports from the intervention GPs, but these have not yet been subject to systematic analyses. However, it is our impression that the GPs had problems restricting their reporting to information about functional abilities and avoiding medical information in the reports, which represents an important aspect in relation to patient confidentiality. This issue was addressed at the workshop, but it seems evident that more instruction, better guidelines, and possibly specific GP training are needed to avoid transfer of medical information that threatens the patient confidentiality. A closer examination of these Functional Assessment Reports represents a potential future study.

Potential application of structured functional assessments in Norwegian general practice

A functional ability model has been discussed within the social security medicine since 1950, and the debate has been reflected in the subsequently introduced disability pension application forms (161). So far, no functional assessment model has been implemented in the Norwegian social insurance system, and most other welfare states have not implemented such methods either. However, one exception may be the Danish assessment method, 'Arbejdsevnemetoden', which seems to be based on a functional ability model (see section 1.3.5). Solli (161) has presented arguments for an implementation of a 'complex functional ability model' in the Norwegian social insurance system.

A new methodology for work ability assessments, sharing some common features with the Danish method, will be implemented in NAV during the winter 2009. The purpose is to provide an early assessment of work possibilities, resources and barriers by involving the sick-listed person. In this context, the structured functional assessment method may provide a feasible way of transferring relevant information from the GPs to the executive officers, who shall make decisions with regards to workplace adjustments, rehabilitation, and social security benefits.

If the structured functional assessment method had been implemented in the Norwegian general practice, we cannot expect that the method would have been welcomed in the same way that it was in the cluster RCT, as this sample represented a selected group of interested GPs. Hence, it may be hypothesised that an implementation rate would be somewhat lower in the Norwegian general practice than in the present study. It is not known if the low GP participation rate in the present study is related to an in general limited interest for social insurance medicine among Norwegian GPs, or if it is related to other factors, for instance heavy workload among the invited GPs.

The method is relatively time-consuming for GPs in busy primary care practices and should therefore be used selectively. It is probably most useful in complex sick leave cases where the GPs identify a need for a more thorough assessment. An electronic version of the forms would probably be more time-efficient. This has already been assessed in a small study in general practice, and despite low participation and low implementation rates, the response from the participating GPs was positive (92).

The GPs' motivation to find time for an expanded consultation and apply the structured method seemed to be partly dependent on whether they received feedback from the local social security officer and/or the employer. Naturally, the GPs were not interested in doing extra work if they doubted that the report information would be utilised or if they received no feedback. During the workshop, the GPs expressed the importance of receiving feedback from local social security officers and/or employers, either on the report content or on consequences of the information and suggestions in the report. All in all, good communication between the involved stakeholders seems to represent an important prerequisite for successful method implementation and for improving the process of return to work.

6. Conclusions

Females, older persons and persons with lower levels of education reported more functional problems than males, younger persons and persons with higher levels of education in The Ullensaker Study 2004. Normative population scores for the Norwegian Function Assessment Scale (NFAS) are presented by gender, age, and length of education. This data may serve as a basis for the development of national population norms. Such norms may be useful in epidemiological research for sample score comparisons or for individual score comparisons in clinical settings and social insurance.

The NFAS demonstrated acceptable validity and reliability as an epidemiological measure of work-related functional ability in a Norwegian population based sample. These findings advocate an application of the NFAS in epidemiological research on work-related functional ability.

Both the original four-point and the new five-point scale version of the NFAS had acceptable data quality and validity, and they may both be used in epidemiological research or in clinical contexts. However, the five-point version had slightly better results for some indices and is therefore recommended for future use. To further test the NFAS, the instrument's responsiveness and test-retest reliability of the five-point version should be assessed. Including a 'not applicable' response option should also be considered.

Implementing structured functional assessments for persons with long-term sick leave in general practice made the GPs capable to assess functional ability in a structured manner. The modest intervention, a one-day workshop with some telephone support, enhanced GP knowledge about functional assessments and patient work factors with sustained effects at the follow-up. Further, the GPs reported higher self-efficacy towards performing functional assessments, but their attitude in relation to functional assessments remained stable.

The intervention GPs prescribed significantly more part-time and less active sick leave than the control GPs, whereas no intervention effect was seen on duration of patient sick leave episodes or on GP prescription of vocational rehabilitation. Part-time sick leave might be important to prevent or reduce the negative effects associated with long-term sick leave. An important aspect in the structured functional assessment method is that it actively involves the sick-listed person and captures their own experience of functional ability. However, caution should be applied to avoid potential adverse effects like information misuse or transfer of information that threatens patient confidentiality.

References

(1) Tellnes G. Sickness certification in general practice: a review. Fam Pract 1989 Mar;6(1):58-65.

(2) Bandura A. Self-efficacy. The Exercise of Control. New York: W.H. Freeman and Company; 1998.

(3) Foucault M. The birth of the clinic: An archaeology of medical perception. New York: Random House; 1975.

(4) Engel GL. The need for a new medical model: a challenge for biomedicine. Science 1977 Apr 8;196(4286):129-36.

(5) Wade DT, Halligan PW. Do biomedical models of illness make for good healthcare systems? BMJ 2004 Dec 11;329(7479):1398-401.

(6) Ihlebaek C, Brage S, Eriksen HR. Health complaints and sickness absence in Norway, 1996-2003. Occup Med (Lond) 2007 Jan;57(1):43-9.

(7) World Health Organization. ICF-International Classification of Functioning, Disability, and Health. Geneva: World Health Organization; 2001.

(8) Cieza A, Ewert T, Ustun TB, Chatterji S, Kostanjsek N, Stucki G. Development of ICF Core Sets for patients with chronic conditions. J Rehabil Med 2004 Jul;(44 Suppl):9-11.

(9) Stucki G, Kostanjsek N, Ustun B, Cieza A. ICF-based classification and measurement of functioning. Eur J Phys Rehabil Med 2008 Sep;44(3):315-28.

(10) Nordenfelt L. Action, ability, and health. Dordrecht: Kluwer Academic Publishers; 2000.

(11) WONCA Classification Committee. Functional Status Measurement in Primary Care. New York/Berlin/Heidelberg: Springer Verlag; 1990.

(12) Natvig B. Musculoskeletal pain. From humdrum nuisance to chronic, disabling pain. [PhD Thesis]. Oslo: University of Oslo; 2001.

(13) Ilmarinen J, Tuomi K. Work ability of aging workers. Scand J Work Environ Health 1992;18 Suppl 2:8-10.

(14) Ilmarinen J. Ageing workers in the European Union – Status and promotion of work ability, employability and employment. Helsinki: Institute of Occupational Health, Ministry of Social Affairs and Health, Ministry of Labour; 1999.

(15) Norwegian Official Reports. [New disability pension allowance and new retirement pension for disability pensioners. NOU 2007:4 (In Norwegian)]. Oslo: Seksjon statens trykking; 2007.

(16) Reiso H. Work ability and sickness absence. A follow-up study in general practice. [PhD thesis]. Oslo: University of Oslo; 2004.

(17) Schult M. [Work ability assessment. What decreases the work ability and how do we assess work ability? (In Swedish)]. Svensk Rehabilitering 2000;(4).

(18) Krohne K, Brage S. [Health care professionals' understanding of function. The project "Functional Assessments- Health Providers' Responsibilities in a More Inclusive Working Life" (In Norwegian)]. Report 1/2007 ed. University of Oslo: Section of Occupational Health and Social Insurance Medicine; 2007.

(19) Engbers M, Veiersted B. [Report from The Expert Conference on Functional Assessments 27. -28. January 2003 (In Norwegian)]. Årg. 4, nr. 4 ed. Oslo: Statens arbeidsmiljøinstitutt; 2003.

(20) Bentsen BG, Natvig B, Winnem M. [Assessment of one's own functional status. COOP-WONCA questionnaire charts in clinical practice and research (In Norwegian)]. Tidsskr Nor Laegeforen 1997 May 10;117(12):1790-3.

(21) van Weel C, König-Zahn C, Touw-Otten FWMM, van Duijn NP, Meyboom-de Jong B. Measuring functional health status with the COOP/Wonca Charts. Groningen: Northern Centre for Health Care Research, University of Groningen, The Netherlands; 1995.

(22) Cocchiarella L, Andersson GBJ. Guides to the evaluation of permanent impairment. United States of America: American Medical Association; 2002.

(23) Kessler R, Stang P, editors. Health and Work Productivity. Making the Business Case for Quality Health Care. Chicago and London: The University of Chicago Press; 2006.

(24) McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N, et al. Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. Health Technol Assess 2001;5(31):1-256.

(25) Streiner DL, Norman GR. Health measurement scales: A practical guide to their development and use. Third ed. Oxford: Oxford University Press; 2003.

(26) Ware JEJr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992 Jun;30(6):473-83.

(27) Ware JEJr, Snow KK, Kosinski M, Gandek B. SF-36® Health Survey Manual and Interpretation Guide. Boston, MA: The Health Institute; 1993.

(28) Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E. A quantitative approach to perceived health status: a validation study. J Epidemiol Community Health 1980 Dec;34(4):281-6.

(29) EuroQoL Group. EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. Health Policy 1990 Dec;16(3):199-208.

(30) Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine 1983 Mar;8(2):141-4.

(31) Doward LC, Spoorenberg A, Cook SA, Whalley D, Helliwell PS, Kay LJ, et al. Development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis. Ann Rheum Dis 2003 Jan;62(1):20-6.

(32) Guyatt GH, Townsend M, Berman LB, Keller JL. A comparison of Likert and visual analogue scales for measuring change in function. J Chronic Dis 1987;40(12):1129-33.

(33) Cook DJ, Guyatt GH, Adachi JD, Epstein RS, Juniper EF, Austin PA, et al. Development and validation of the mini-osteoporosis quality of life questionnaire (OQLQ) in osteoporotic women with back pain due to vertebral fractures. Osteoporosis Quality of Life Study Group. Osteoporos Int 1999;10(3):207-13.

(34) Juniper EF, O'Byrne PM, Guyatt GH, Ferrie PJ, King DR. Development and validation of a questionnaire to measure asthma control. Eur Respir J 1999 Oct;14(4):902-7.

(35) Preston CC, Colman AM. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychol (Amst) 2000 Mar;104(1):1-15.

(36) Boyce A, In: Cox EP. The Optimal Number of Response Alternatives for a Scale: A Review. J Marketing Research 1980;17:407-22.

(37) Cox EP. The Optimal Number of Response Alternatives for a Scale: A Review. J Marketing Research 1980;17:407-22.

(38) Loge JH, Kaasa S, Hjermstad MJ, Kvien TK. Translation and performance of the Norwegian SF-36 Health Survey in patients with rheumatoid arthritis. I. Data quality, scaling assumptions, reliability, and construct validity. J Clin Epidemiol 1998 Nov;51(11):1069-76.

(39) Brazier JE, Harper R, Jones NM, O'Cathain A, Thomas KJ, Usherwood T, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. BMJ 1992 Jul 18;305:160-4.

(40) Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. BMJ 2002 Jun 15;324(7351):1417.

(41) Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. J R Coll Gen Pract 1985 Apr;35(273):185-8.

(42) Spilker B. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. Philadelphia: Lippincott-Raven; 1996.

(43) Brooks R, Rabin R, De Charro F, editors. The measurement and valuation of health status using EQ-5D: a European perspective. Evidence from the EuroQol BIOMED research programme. Dordrecht: Kluwer academic publ; 2003.

(44) Dolan P. Modeling valuations for EuroQol health states. Med Care 1997 Nov;35(11):1095-108.

(45) Rabin R, de CF. EQ-5D: a measure of health status from the EuroQol Group. Ann Med 2001 Jul;33(5):337-43.

(46) Essink-Bot ML, Krabbe PF, Bonsel GJ, Aaronson NK. An empirical comparison of four generic health status measures. The Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. Med Care 1997 May;35(5):522-37.

(47) Nelson E, Wasson J, Kirk J, Keller A, Clark D, Dietrich A, et al. Assessment of function in routine clinical practice: description of the COOP Chart method and preliminary findings. J Chronic Dis 1987;40 Suppl 1:55S-69S.

(48) Scholten JHG, van Weel C. Functional status assessment in family practice: the Dartmouth COOP Functional Health Assessment Charts/WONCA. Lelystad: Meditekst; 1992.

(49) Kinnersley P, Peters T, Stott N. Measuring functional health status in primary care using the COOP-WONCA charts: acceptability, range of scores, construct vaildity, reliability and sensitivity to change. Br J Gen Pract 1994;44:545-9.

(50) Bentsen BG, Natvig B, Winnem M. Questions you didn't ask? COOP/WONCA Charts in clinical work and research. World Organization of Colleges, Academies and Academic Associations of General Practitioners/Family Physicists. Fam Pract 1999 Apr;16(2):190-5.

(51) Lindegaard PM, Bentzen N, Christiansen T. Reliability of the COOP/WONCA charts. Test-retest completed by patients presenting psychosocial health problems to their general practitioner. Scand J Prim Health Care 1999 Sep;17(3):145-8.

(52) Stavem K, Jodalen H. Reliability and validity of the COOP/WONCA health status measure in patients with chronic obstructive pulmonary disease. Quality of Life Research 2002 Sep;11(6):527-33.

(53) Bruusgaard D, Nessioy I, Rutle O, Furuseth K, Natvig B. Measuring functional status in a population survey. The Dartmouth COOP functional health assessment charts/WONCA used in an epidemiological study. Fam Pract 1993 Jun;10(2):212-8.

(54) Social Security Office. A guide to Incapacity Benefit - The Personal Capability Assessment. Leeds: DWP Communications; 2002.

(55) Swales K, Craig P. Evaluation of the Incapacity Benefit Medical Test: In-house report 26. London: Social Research Branch, Analytical Services Division, Department of Social Security; 1997.

(56) Tuomi K, Ilmarinen J, Jahkola A, Katajarinne L, Tulkki A. Work Ability Index. 2nd revised ed. Helsinki: Finnish Institute of Occupational Health; 1998.

(57) Nygard CH, Eskelinen L, Suvanto S, Tuomi K, Ilmarinen J. Associations between functional capacity and work ability among elderly municipal employees. Scand J Work Environ Health 1991;17 Suppl 1:122-7.

(58) European Commission. Definitions of disability in Europe. A comparative analysis. [Online] 2002 [cited 2008 Aug 13];Available from: URL: http://europa.eu.int/comm/employment_social/index/complete_report_en.pdf

(59) Arbeidsmarkedsstyrelsen. *Arbejdsevnemetode [Work Ability Method]*. København: 2002.

(60) Norwegian Official Reports. [Sick leave and disability pensioning. An inclusive working life. NOU 2000:27 (In Norwegian)]. Oslo: Seksjon statens trykking; 2000.

(61) St.prp.nr.1 Tillegg nr.1 (2001-2002). [The tripartite agreement on a more inclusive workplace (In Norwegian)]. Oslo: 2001.

(62) Brage S, Fleten N, Knudsrod OG, Reiso H, Ryen A. [Norwegian Functional Scale--a new instrument in sickness certification and disability assessments (In Norwegian)]. Tidsskr Nor Laegeforen 2004 Oct 7;124(19):2472-4.

(63) World Health Organization. ICF Checklist. Version 2.1a, Clinical Form for International Classification of Functioning, Disability and Health. [Online] 2007 [cited 2007 Nov 15];Available from: URL: http://www.e-bility.gr/eutexnos/Includes/icf-checklist.pdf

(64) Timpka T, Hensing G, Alexanderson K. Dilemmas in sickness certification among Swedish physicians. Eur J Pub Health 1995;(5):215-9.

(65) Krakau I. Severity of illness and diagnoses in a Swedish general practice population. Fam Pract 1991 Mar;8(1):28-31.

(66) Andersen EJ. [Functional assessments causing trouble? (In Norwegian]. Tidsskr Nor Laegeforen 2003;123(20):2928.

(67) Krohne K, Brage S. New rules meet established sickness certification practice. A focus group study on the introduction of functional assessments in Norwegian primary care. Scand J Prim Health Care 2007;25(3):172-7.

(68) Health and Benefits Division. Transformation of the Personal Capability Assessment. Report of the Physical Function and Mental Health Technical Working Groups. London: Department for Work and Pensions; 2006.

(69) Anonymous. A guide to the Employment and Support Allowance – The Work Capability Assessment. [Online] 2008 [cited 2008 Oct 17];Available from: URL: http://www.jobcentreplus.gov.uk/JCP/stellent/groups/jcp/documents/websitecontent/dev_015887.pdf

(70) Arbejdsmarkedsstyrelsen. [Work Ability Method. Dialogue and co-operation (In Danish)]. København: Arbejdsmarkedsstyrelsen, Bæskjeftigelsesministeriet; 2002.

(71) Socialministeriet. [Work Ability Method. Method for description, development and assessment of work ability (In Danish)]. Køpenhavn: Socialministeriet; 2001.

(72) Center for Aktiv Beskæftigelsesindsats. [The Work Ability Method (In Danish)]. [Online] 2008 [cited 2008 Oct 17];Available from: URL: http://www.fleksjob2.dk/lovgivning/AktivBeskaeftIndsats/Arbejdsevne/Arbejdsevne.htm

(73) Høgelund J. In search of effective disability policy. Comparing the developments and outcomes of the Dutch and Danish disability policies. Amsterdam: Amsterdam University Press; 2003.

(74) Engbers M, Furulund P. [Functional assessments at the workplace, a tool for special assessments by the company health service. A test of the Demand and Function Scheme in the field (In Norwegian)]. [Online] 2006 January 31 [cited 2008 Oct 16];Report to NHO, the Confederation of Norwegian Enterprise(Project no. S-2387)Available from: URL: http://www.funksjonsvurdering-bht.no/rapporter_lenker.php

(75) Aben I, Van Lierop B, Janssen N, Nijhuis F. CBBS, the new Dutch way to determine the level of work incapacity of disabled people. European Journal of Public Health 2003;13(Suppl. 2):84.

(76) Tellnes G. Sickness Certification - an Epidemiological Study Related to Community Medicine and General Practice [PhD thesis]. Oslo: University of Oslo; 1990.

(77) The Ministry of Labour and Social Inclusion. [The National Insurance Act (In Norwegian)]. [Online] 2008 June 27 [cited 2008 Oct 28];Available from: URL: http://www.lovdata.no/all/nl-19970228-019.html

(78) Norwegian Labour and Welfare Administration. [GPs' sickness certification practices II: New rules and The GP Education Programme. RTV-Report 06/2006 (In Norwegian)]. Oslo: 2006.

(79) Norwegian Labour and Welfare Administration. [Persons receiving social security benefits and other arrangements administrated by the Norwegian Labour and Welfare Administration. 1997-2006 (In Norwegian)]. [Online] 2007 November 6 [cited 2008 May 2];Available from: URL: http://www.nav.no/805365427.cms

(80) Daehli B, Bjerkedal T. [Physicians practicing for the legal requirements for eligibility for disability pensions (In Norwegian)]. Tidsskr Nor Laegeforen 1991 Apr 10;111(9):1144-6.

(81) Gulbrandsen P, Hofoss D, Nylenna M, Saltyte-Benth J, Aasland OG. General practitioners' relationship to sickness certification. Scand J Prim Health Care 2007 Mar;25(1):20-6.

(82) Wahlstrom R, Alexanderson K. Swedish Council on Technology Assessment in Health Care (SBU). Chapter 11. Physicians' sick-listing practices. Scand J Public Health Suppl 2004;63:222-55.

(83) Rogers EM. Diffusion of innovation. 4th ed. New York: The free press; 1995.

(84) Hjortdahl P. [Why and how should the clinical behavior of physicians be changed? (In Norwegian)]. Tidsskr Nor Laegeforen 1994 Aug 30;114(20):2372-3.

(85) Lomas J. Teaching old (and not so old) docs new tricks: effective ways to implement research findings. In: Dunn E, Norton P, Stewart M, Tudiver F, Bass M, editors. Disseminating research/changing practice.Thousand Oaks: Sage; 1994. p. 1-18.

(86) Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 1977;84(2):117-25.

(87) O'Brien MA, Freemantle N, Oxman AD, Wolf F, Davis DA, Herrin J. Continuing education meetings and workshops: effects on professional practice and health care outcomes. Cochrane Database Syst Rev 2001;(2):CD003030.

(88) Grimshaw JM, Shirran L, Thomas R, Mowatt G, Fraser C, Bero L, et al. Changing provider behavior: an overview of systematic reviews of interventions. Med Care 2001 Aug;39(8), Suppl 2:II-2-II-45.

(89) Bero LA, Grilli R, Grimshaw JM, Harvey E, Oxman AD, Thomson MA. Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. The Cochrane Effective Practice and Organization of Care Review Group. BMJ 1998 Aug 15;317(7156):465-8.

(90) Kendrick T, Burns T, Freeling P. Randomised controlled trial of teaching general practitioners to carry out structured assessments of their long term mentally ill patients. BMJ 1995 Jul 8;311(6997):93-8.

(91) Scheel IB. The effects of promoting active sick leave for back pain patients. [PhD thesis]. Oslo: University of Oslo; 2004.

(92) Brage S, Østerås N, Krohne K, Steiran P. [Functional assessments in relation to long-term sick leave and the decision on long-term social benefits (In Norwegian)]. University of Oslo: Section of Occupational Health and Social Insurance Medicine; 2008.

(93) Statistics Norway. StatBank Norway. www ssb no 2006 March 1 [cited 2006 Oct 10];Available from: URL: www.ssb.no

(94) Osteras N, Brage S, Garratt A, Benth JS, Natvig B, Gulbrandsen P. Functional ability in a population: normative survey data and reliability for the ICF based Norwegian Function Assessment Scale. BMC Public Health 2007;7:278.

(95) Osteras N, Gulbrandsen P, Garratt A, Benth JS, Dahl FA, Natvig B, et al. A randomised comparison of a four- and a five-point scale version of the Norwegian Function Assessment Scale. Health Qual Life Outcomes 2008 Feb 15;6(1):14.

(96) Reiso H, Nygard JF, Brage S, Gulbrandsen P, Tellnes G. Work ability assessed by patients and their GPs in new episodes of sickness certification. Fam Pract 2000 Apr;17(2):139-44.

(97) Goldberg DP. Manual of the General Health Questionnaire. Windsor: 1978.

(98) Kuorinka I, Jonsson B, Kilbom A, Vinterberg H, Biering-Sorensen F, Andersson G, et al. Standardised Nordic questionnaires for the analysis of musculoskeletal symptoms. Appl Ergon 1987 Sep;18(3):233-7.

(99) Fleiss JK. Statistical Methods for Rates and Proportions, 2nd edn. New York: Wiley; 1981.

(100) Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.

(101) Kaasa S, Bjordal K, Aaronson N, Moum T, Wist E, Hagen S, et al. The EORTC core quality of life questionnaire (QLQ-C30): validity and reliability when analysed with patients treated with palliative radiotherapy. Eur J Cancer 1995 Dec;31A(13-14):2260-3.

(102) Norwegian Labour and Welfare Administration. [Principal number report, per 4. quarter 2005 (In Norwegian]. [Online] 2006 February 10 [cited 2006 Dec 21];Available from: URL: http://www.nav.no/binary/1073748592/file

(103) Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.

(104) Norwegian Insurance Administration. [Yearbook of National Insurance Statistics 2003 (In Norwegian)]. Oslo: NIA; 2003.

(105) Arbuckle J, Wothke W. Amos 4.0 User's guide. Chicago: SmallWaters Corporation; 1999.

(106) Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974;19(6):716-23.

(107) Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. Hoboken, N.J: Wiley-Interscience; 2004.

(108) Massof RW. Likert and Guttman scaling of visual function rating scale questionnaires. Ophthalmic Epidemiol 2004 Dec;11(5):381-99.

(109) Hays RD, Ellickson PL. Guttman scale analysis of longitudinal data: a methodology and drug use applications. Int J Addict 1990;25(11A):1341-52.

(110) Sackett DL. Bias in analytic research. J Chronic Dis 1979;32(1-2):51-63.

(111) Kleinbaum DG, Kupper LL, Morgenstern H. Validity of Epidemiologic Research. In: Kleinbaum DG, Kupper LL, Morgenstern H, editors. Epidemiologic research: Principles and quantitative methods.New York: Van Nostrand Reinhold; 1982.

(112) Rothman KJ. Epidemiology: An Introduction. Oxford: Oxford University Press; 2002.

(113) Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. Health Technology Assessment 1998;2(14).

(114) Deyo RA. Measuring the functional status of patients with low back pain. Arch Phys Med Rehabil 1988 Dec;69(12):1044-53.

(115) Woodward M. Epidemiology: Study design and data analysis. Boca Raton: Chapman & Hall/CRC; 2005.

(116) Holm I, Friis A, Storheim K, Brox JI. Measuring self-reported functional status and pain in patients with chronic low back pain by postal questionnaires: a reliability study. Spine 2003 Apr 15;28(8):828-33.

(117) Stratford PW, Kennedy D, Pagura SM, Gollish JD. The relationship between self-report and performance-related measures: questioning the content validity of timed tests. Arthritis Rheum 2003 Aug 15;49(4):535-40.

(118) Fordyce WE, Lansky D, Calsyn DA, Shelton JL, Stolov WC, Rock DL. Pain measurement and pain behavior. Pain 1984 Jan;18(1):53-69.

(119) Haugland S. School related stress, physical activity, and health complaints in adolescence. [PhD thesis]. Bergen: University of Bergen; 2001.

(120) Simmonds MJ, Olson SL, Jones S, Hussein T, Lee CE, Novy D, et al. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. Spine 1998 Nov 15;23(22):2412-21.

(121) Gandhi R, Tsvetkov D, Davey JR, Syed KA, Mahomed NN. Relationship between self-reported and performance-based tests in a hip and knee joint replacement population. Clin Rheumatol 2008 Oct 8;[Epub ahead of print].

(122) Rallon CR, Chen CC. Relationship between performance-based and self-reported assessment of hand function. Am J Occup Ther 2008 Sep;62(5):574-9.

(123) Loge JH, Kaasa S. Short form 36 (SF-36) health survey: normative data from the general Norwegian population. Scand J Soc Med 1998 Dec;26(4):250-8.

(124) Sullivan M, Karlsson J, Ware JE, Jr. The Swedish SF-36 Health Survey--I. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. Soc Sci Med 1995 Nov;41(10):1349-58.

(125) Jenkinson C, Stewart-Brown S, Petersen S, Paice C. Assessment of the SF-36 version 2 in the United Kingdom. J Epidemiol Community Health 1999 Jan;53(1):46-50.

(126) Sogaard AJ, Selmer R, Bjertness E, Thelle D. The Oslo Health Study: The impact of self-selection in a large, population-based survey. Int J Equity Health 2004 May 6;3(1):3.

(127) Blake C, Codd MB, O'Meara YM. The Short Form 36 (SF-36) Health Survey: normative data for the Irish population. Ir J Med Sci 2000 Jul;169(3):195-200.

(128) Hopman WM, Towheed T, Anastassiades T, Tenenhouse A, Poliquin S, Berger C, et al. Canadian normative data for the SF-36 health survey. Canadian Multicentre Osteoporosis Study Research Group. CMAJ 2000 Aug 8;163(3):265-71.

(129) Martikainen P, Laaksonen M, Piha K, Lallukka T. Does survey non-response bias the association between occupational social class and health? Scand J Public Health 2007;35(2):212-5.

(130) van den AM, Buntinx F, Metsemakers JF, Knottnerus JA. Morbidity in responders and non-responders in a register-based population survey. Fam Pract 1998 Jun;15(3):261-3.

(131) Korkeila K, Suominen S, Ahvenainen J, Ojanlatva A, Rautava P, Helenius H, et al. Non-response and related factors in a nation-wide health survey. Eur J Epidemiol 2001;17(11):991-9.

(132) Bostrom G, Hallqvist J, Haglund BJ, Romelsjo A, Svanstrom L, Diderichsen F. Socioeconomic differences in smoking in an urban Swedish population. The bias introduced by non-participation in a mailed questionnaire. Scand J Soc Med 1993 Jun;21(2):77-82.

(133) Hoffman SC, Burke AE, Helzlsouer KJ, Comstock GW. Controlled trial of the effect of length, incentives, and follow-up techniques on response to a mailed questionnaire. Am J Epidemiol 1998 Nov 15;148(10):1007-11.

(134) Hartman JM, Forsen JW, Jr., Wallace MS, Neely JG. Tutorials in clinical research: part IV: recognizing and controlling bias. Laryngoscope 2002 Jan;112(1):23-31.

(135) Dahl FA, Grotle M, Saltyte BJ, Natvig B. Data splitting as a countermeasure against hypothesis fishing: with a case study of predictors for low back pain. Eur J Epidemiol 2008;23(4):237-42.

(136) van den Brink CL, Tijhuis M, Kalmijn S, Klazinga NS, Nissinen A, Giampaoli S, et al. Self-reported disability and its association with performance-based limitation in elderly men: a comparison of three European countries. J Am Geriatr Soc 2003 Jun;51(6):782-8.

(137) Avis NE, Smith KW. Conceptual and methodological issues in selecting and developing quality of life measures. In: Fitzpatrick R, editor, editors. Advances in medical sociology.London: JAI Press Inc.; 2006. p. 255-80.

(138) Holmstrom E, Moritz U. Low back pain--correspondence between questionnaire, interview and clinical examination. Scand J Rehabil Med 1991;23(3):119-25.

(139) Goldberg D, Williams P. A user's guide to the general health questionnaire. Berkshire: NFER-Nelson; 1988.

(140) Malt UF, Mogstad TE, Refnin IB. [Goldberg's General Health Questionnaire (In Norwegian)]. Tidsskr Nor Laegeforen 1989 May 10;109(13):1391-4.

(141) Malt UF. The validity of the General Health Questionnaire in a sample of accidentally injured adults. Acta Psychiatr Scand Suppl 1989;355:103-12.

(142) Haldorsen EM, Indahl A, Ursin H. Patients with low back pain not returning to work. A 12-month follow-up study. Spine 1998 Jun 1;23(11):1202-7.

(143) Reiso H, Nygard JF, Brage S, Gulbrandsen P, Tellnes G. Work ability and duration of certified sickness absence. Scand J Public Health 2001 Sep;29(3):218-25.

(144) Reiso H, Nygard JF, Jorgensen GS, Holanger R, Soldal D, Bruusgaard D. Back to work: predictors of return to work among patients with back disorders certified as sick: a two-year follow-up study. Spine 2003 Jul 1;28(13):1468-73.

(145) Allebeck P, Mastekaasa A. Swedish Council on Technology Assessment in Health Care (SBU). Chapter 3. Causes of sickness absence: research approaches and explanatory models. Scand J Public Health Suppl 2004;63:36-43.

(146) Bombardier C. Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. Spine 2000 Dec 15;25(24):3100-3.

(147) Terwee CB, van der Slikke RM, van Lummel RC, Benink RJ, Meijers WG, de Vet HC. Self-reported physical functioning was more influenced by pain than performance-based physical functioning in knee-osteoarthritis patients. J Clin Epidemiol 2006 Jul;59(7):724-31.

(148) Myers AM, Holliday PJ, Harvey KA, Hutchinson KS. Functional performance measures: are they superior to self-assessments? J Gerontol 1993 Sep;48(5):M196-M206.

(149) Swedish Official Reports. [Work ability? A review of assessment methods in Sweden and other countries. SOU 2008:66 (In Swedish)]. Stockholm: 2008.

(150) Cook TD, Campbell DT. Quasi-Experimentation: Design & Analysis Issues for Field Settings. Boston: Houghton Mifflin; 1979.

(151) Rokstad K, Fugelli P. How to succeed in a multipractice study. Fam Med 1993 Jul;25(7):461-4.

(152) Gulbrandsen P. Social context in general practice. Disclosure, recognition, and management. Thesis. University of Oslo, Institute of General Practice and Community Medicine, 1998.

(153) Garratt AM, Ruta DA, Abdalla MI, Buckingham JK, Russell IT. The SF-36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? BMJ 1993 May 29;306(6890):1440-4.

(154) Ware JEJr, Kosinski M, Gandek B, Aaronson NK, Apolone G, Bech P, et al. The factor structure of the SF-36 Health Survey in 10 countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998 Nov;51(11):1159-65.

(155) Vingard E, Alexanderson K, Norlund A. Swedish Council on Technology Assessment in Health Care (SBU). Chapter 9. Consequences of being on sick leave. Scand J Public Health Suppl 2004;63:207-15.

(156) Waddell G, Burton K, Aylward M. Work and common health problems. J Insur Med 2007;39(2):109-20.

(157) Waddell G, Aylward M, Sawney P. Back pain, incapacity for work and social security benefits. An international literature review and analysis. London: Royal Society of Medicine Press; 2002.

(158) Ockander M, Timpka T. A female lay perspective on the establishment of long-term sickness absence. International Journal of Social Welfare 2001 Jan;10(1):74-9.

(159) Nordqvist C, Holmqvist C, Alexanderson K. Views of laypersons on the role employers play in return to work when sick-listed. J Occup Rehabil 2003 Mar;13(1):11-20.

(160) Sieurin L, Josephson M, Vingard E. Positive and negative consequences of sick leave for the individual, with special focus on part-time sick leave. Scand J Public Health 2009 Jan;37(1):50-6.

(161) Solli HM. Rettferdighet og objektivitet i trygdemedisinske uførhetsvurderinger. En etisk og vitenskapsfilosofisk analyse av tre uførhetsmodeller i et historisk perspektiv (In Norwegian). [PhD thesis]. Oslo: University of Oslo; 2007.

I

Functional ability in a population: normative survey data and reliability for the ICF based Norwegian Function Assessment Scale. Nina Østerås, Søren Brage, Andrew Garratt, Jurate Saltyte Benth, Bård Natvig and Pål Gulbrandsen. BMC Public Health 2007, 7:278

doi:10.1186/1471-2458-7-278

This article is available from: http://www.biomedcentral.com/1471-2458/7/278

BMC Public Health

BioMed Central

Research article

Open Access

Functional ability in a population: normative survey data and reliability for the ICF based Norwegian Function Assessment Scale

Nina Østerås*[1], Søren Brage[1], Andrew Garratt[2], Jurate Saltyte Benth[3,4], Bård Natvig[1] and Pål Gulbrandsen[3,4]

Address: [1]Section of Occupational Health and Social Insurance Medicine, Institute of General Practice and Community Health, Faculty of Medicine, University of Oslo, Norway, [2]Institute of Health Management and Health Economics, University of Oslo, Norway, [3]Helse Øst Health Services Research Centre, Akershus University Hospital, Norway and [4]Faculty of Medicine, University of Oslo, Norway

Email: Nina Østerås* - nina.osteras@medisin.uio.no; Søren Brage - soren.brage@medisin.uio.no; Andrew Garratt - andrew.garratt@kunnskapssenteret.no; Jurate Saltyte Benth - jurate@ahus.no; Bård Natvig - bard.natvig@medisin.uio.no; Pål Gulbrandsen - pal.gulbrandsen@ahus.no

* Corresponding author

Published: 3 October 2007

BMC Public Health 2007, **7**:278 doi:10.1186/1471-2458-7-278

Received: 1 November 2006
Accepted: 3 October 2007

This article is available from: http://www.biomedcentral.com/1471-2458/7/278

Abstract

Background: The increasing focus on functional ability assessments in relation to sickness absence necessitates the measurement of population functional levels. This study assessed the reliability of the Norwegian Function Assessment Scale (NFAS) and presents normative population data.

Methods: All inhabitants in seven birth cohorts in Ullensaker municipality in 2004 were approached by means of a postal questionnaire. The NFAS was included as part of The Ullensaker Study 2004. The instrument comprises 39 items derived from the activities/participation component in the International Classification for Functioning, Disabilities and Health (ICF). Based on the results of principal component analysis, these items comprise seven domains. Non-parametric tests for independent samples were used to compare subgroups. Internal consistency was assessed by Cronbach's alpha. Two-week test-retest reliability was assessed by total proportions of agreement, weighted kappa, and intraclass correlation coefficient (ICC).

Results: The response rate was 54% (1620 persons) and 75.4% (101 persons) for the retest. Items had low levels of missing data. Test-retest reliability was acceptable with high proportions of absolute agreement; kappa and ICC values ranged from 0.38 to 0.83 and 0.79 to 0.83, respectively. No difficulty on all 39 functional activities was reported by 33.1% of respondents. Females, older persons and persons with lower levels of education reported more functional problems than their respective counterparts ($p < 0.05$). The age gradient was most evident for three of the physical domains. For females aged 24–56 and males aged 44–76, a clear education gradient was present for three of the physical domains and one mental domain after adjusting for age and gender.

Conclusion: This study presents population based normative data on functional ability, as measured by the NFAS. These data will serve as basis for the development of national population norms and are necessary for score interpretation. Data quality and test-retest reliability of the NFAS were acceptable.

Background

Longitudinal trends in sickness absence and disability pensions rates in several European countries, including Norway, show that increasing proportions of the population have levels of work ability that are too low to meet work demands [1]. To meet this challenge, European social security schemes increasingly emphasize the individual's resources and functional abilities rather than health deficits and restrictions. The Norwegian Insurance Scheme has introduced functional ability assessments in sickness certification forms [2]. In this context, the new classification for functioning, disabilities and health (ICF) has received attention through its consistent conceptual framework for defining functional ability [3].

It is commonly found that the level of functioning tends to be poorer with increasing age and in lower social classes [4]. Eurostat surveys conducted within the European Union, have estimated the prevalence rates for moderate and severe disability in the working-age population to be 10.0% and 4.5% respectively [4]. These figures were based on self-assessed restrictions in daily activities – moderate or severe – and stricter definitions of functional limitations would give lower prevalence figures.

National and international population surveys have frequently used well-established health status instruments such as the Nottingham Health Profile [5] and the Short Form 36-item (SF-36) Health Survey [6] to assess function. Such questionnaires often have multiple aims and include several scales to measure function and quality of life. A broad array of scales might be relevant for clinical and epidemiological investigations, but are less useful in social security. To reintegrate employees in working life, there is a need for discriminating instruments that can aid the medical assessors, case managers, and labour experts in their decisions as to who should receive which types of benefits and support. Instruments based on self-report have been developed in the UK, the Netherlands, and Finland [7,8]. Self-reports of health conditions, abilities, and skills are also important approaches in the expanding research field on the relationships between health and work productivity [9]. In the WHO Health and Work Performance Questionnaire, functional status is reported, although on a more general level [10].

The Norwegian Function Assessment Scale (NFAS) is an instrument for self-report that was developed by an expert group in social insurance in 2000. It was developed to assess the need for rehabilitation, adjustment of work demands among sick-listed persons as well as the rights to social security benefits [11]. ICF was selected as a basis for facilitating multidisciplinary work and understanding, and the usage of generally accepted definition of concepts. All categories from the activities/participation component in ICF were considered, and categories not relevant for the assessment of work-related functional abilities were removed. After this process, 39 categories remained which were rephrased into questions with four response alternatives. Four response alternatives were used in preference to the five within the ICF, because fewer alternatives make the scale easier to use in assessment procedures.

The first version of the NFAS was tested for construct and convergent/divergent validity against SF-36 and the Dartmouth COOP Functional Health Assessment Charts/WONCA(COOP/WONCA), and for utility in a random sample of 386 persons sick-listed for six weeks in eight different geographical areas [11]. Based on a principal component analysis of this data, the 39 categories were regrouped into seven functional domains. Individual assessment of these domains facilitated the design of work place adjustments, and strengthened the communication between the sick-listed person and the case manager in the National Insurance Administration. Recently, the NFAS has been utilized in a study of 89 disability pensioners, to predict belief in return-to-work [12].

The final version of NFAS had good construct validity [11], but its reliability has not yet been thoroughly evaluated. The level of functional ability has yet to be assessed in the general population. This will provide important normative data necessary for score interpretation. Validity of a four- and a five-point scale version of the NFAS in a population will be reported elsewhere. The purpose of this study was to obtain normative data on the NFAS as part of The Ullensaker Study 2004, and to examine the test-retest reliability of the scale.

Methods

Study setting and sample

Ullensaker is a rural community which had 23,700 inhabitants in 2004. There are no major differences between the population of Ullensaker and the population of Norway with respect to demographic characteristics [13]. In 2004, postal questionnaires, which included the NFAS along with questions relating to musculoskeletal pain, were sent to all inhabitants in Ullensaker municipality in the birth cohorts 1918–20, 1928–30, 1938–40, 1948–50, 1958–60, 1968–70 and 1978–80. A randomized half of these inhabitants received the four-point version of NFAS and were included in this study. Reminders were sent at eight weeks. Information on the residential locations was given by the Population Register.

The Regional Committee for Medical Research Ethics and The Norwegian Data Inspectorate approved the study.

Test-retest reliability

For purposes of assessing test-retest reliability, the first 30 returning a questionnaire within each of the five youngest birth cohorts were asked to complete the NFAS again at two weeks. The two oldest birth cohorts were not included because the persons are outside the normal working age in Norway of 16 to 67 years. Individuals reporting no difficulty on all NFAS items were not invited in the retest since possible changes could only be in one direction.

The Norwegian Function Assessment Scale (NFAS)

The NFAS [11] was included in The Ullensaker Study 2004 to obtain self-reported levels of ICF based functional ability. The 39 items are relevant for assessing physical and mental functioning in working life, some relating to activities of daily living. The NFAS starts with the question "Have you had difficulty doing the following activities during the last week?" and respondents self-report 39 activities using a four-point scale from 1–4: no difficulty, some difficulty, much difficulty, could not do it. A low score indicates good functional ability.

Based on the results of principal component analysis from the previous study with sick-listed persons [11], the items comprise seven domains: Walking/standing (7 items), Holding/picking up things (8 items), Lifting/carrying (6 items), Sitting (3 items), Managing (7 items), Cooperation/communication (6 items), Senses (2 items). These domains have evidence for validity in sick-listed persons [11]. The main application of the NFAS is likely to be social insurance. Hence it was decided to keep the domains from the earlier study with sick-listed persons [11]. It should, however, be anticipated that principal component analysis based on data from the general population in Ullensaker will yield somewhat different results. Domain scores are calculated by adding the item scores and dividing by the number of items completed. The NFAS total scores are calculated by adding all 39 item scores and dividing by the number of items completed. Thus, missing values were ignored.

Demographic data about the education level was included in the questionnaire with the response categories of lower secondary school, upper secondary school (technical), upper secondary school (preparatory), university 1–4 years, university >4 years. Education level was then categorized into three groups: ≤ 9 years, 10 to 12 years and ≥13 years.

Statistical analyses

Internal consistency was assessed by Cronbach's alpha. Test-retest reliability was assessed by calculating total proportions of agreement, weighted kappa [14], and intraclass correlation coefficient (ICC) (two-way mixed model with the measure of absolute agreement). Since data are categorical, non-parametric tests for independent samples were used to compare subgroups.

Results

Of the 3000 questionnaires posted, 1620 (54.0%) were returned. Compared to respondents, non-respondents were more likely to be male ($p < 0.001$) and young or very old (Table 1). Of the respondents, 18.5%, 47.5% and 34.1% reported ≤ 9 years, 10 to 12 years and ≥13 years of education respectively.

The mean level of missing data for the 39 NFAS items was 3.3% and 78.5% had no missing data. For the great majority of items, missing values ranged from 1.9 – 4.6%. Holding and turning a steering wheel (5.3%), driving a car (6.1%), working in groups (9.0%), and guiding others in their activities (9.3%) had higher missing values. There was a significant increase of missing values with age ($p < 0.001$).

Item responses were skewed towards no difficulty; range 63.5 – 96.8%. The percentage of respondents reporting no difficulty for all 39 items was 33.1%. The items going up and down stairs, engaging in your leisure activities, pushing and pulling with your arms, cleaning your house, staying alert and being able to concentrate, managing everyday stress and strains, managing to take criticism, managing to control your anger and aggression, and remembering things, represent functional activities in which more than 20% of the population reported difficulties.

Cronbach's alpha ranged from 0.67 (Sitting) to 0.91 (Walking/standing) for domains and was 0.95 for the total scores. Five of seven domains exceeded the 0.70 reliability standard for use in groups [15], the remaining two just failing to meet this criterion.

Table 1: Response rates by age and gender (N = 1620)

	Number included (%)	Response rate (%)
Females	905 (55.9)	60.0
Males	715 (44.1)	48.0
Age:		
24–26	150 (9.3)	33.3
34–36	429 (26.5)	49.9
44–46	301 (18.6)	54.2
54–56	358 (22.1)	68.4
64–66	219 (13.5)	66.2
74–76	132 (8.1)	66.8
84–86	31 (1.9)	37.8

Test-retest

Retest questionnaires were returned by 101 of the 134 (75.4%) individuals sent a second questionnaire. Most persons in the youngest cohort reported no difficulty for all questions, resulting in fewer candidates in this cohort (n = 17). The respondents were significantly older ($p < 0.05$) than the non-respondents, but were otherwise comparable. With the exception of four items – writing, which showed a deterioration ($p < 0.01$) in function, and managing to take criticism, managing to control your aggression and anger, and remembering things, which showed an improvement ($p = 0.01$) in function – there were no score differences between test and retest. The proportion scoring exactly the same on both occasions (total proportions of agreement) was high, ranging from 0.68 – 0.97. Weighted kappa values ranged from 0.38 (fair agreement) to 0.83 (almost perfect agreement) [16] (Table 2). The weighted kappa values for single items showed large variability, but the values for six of the seven domains were above 0.61, indicating good agreement. ICC values ranged from 0.79 (substantial) to 0.88 (almost perfect) [16] (Table 2).

Table 2: Weighted kappa and ICC in a test-retest study using NFAS(N = 101)

No.	N	Weighted kappa	ICC[a]
Walking/standing		**0.66**	**0.85**
1	99	0.47	
2	100	0.60	
3	99	0.70	
4	98	0.66	
5	100	0.65	
6	100	0.64	
7	101	0.76	
Holding/picking up things		**0.67**	**0.87**
8	100	0.76	
9	99	0.40	
10	100	0.83	
11	100	0.71	
12	101	0.72	
13	99	0.51	
14	99	0.57	
15	101	0.70	
Lifting/carrying		**0.65**	**0.82**
16	97	0.58	
17	98	0.69	
18	101	0.56	
19	99	0.64	
20	100	0.76	
21	99	0.52	
Sitting		**0.61**	**0.79**
22	100	0.63	
23	101	0.64	
24	99	0.58	
Managing		**0.57**	**0.79**
25	100	0.67	
26	96	0.75	
27	97	0.48	
28	99	0.57	
29	101	0.56	
30	97	0.51	
31	99	0.38	
Cooperation/communication		**0.68**	**0.88**
32	98	0.64	
33	100	0.61	
34	100	0.60	
35	101	0.74	
36	101	0.64	
37	101	0.47	
Senses		**0.68**	**0.80**
38	100	0.42	
39	101	0.82	

[a] ICC = Intraclass correlation coefficient (two-way mixed with absolute agreement)

Gender

Item and domain scores ranged from 1.04 to 1.42 and from 1.05 to 1.25 respectively (Table 3). Males reported significantly better functional ability than females on 33 items. With the exception of the Cooperation/communication domain, domain and total scores were significantly better for males than females.

Age

Domain and total scores for males and females within different age groups are given in Table 4. With the exception of females in the age group 54–56, the total scores increased gradually with age ($p < 0.001$). With the exception of the Senses domain for males, there is a large deterioration in reported functional ability from second oldest to the oldest age group. When the oldest age cohorts of females and males were excluded, the difference in scores for the Sitting domain became insignificant. Domain scores for three of the physical domains, Walking/standing, Holding/picking up things and Lifting/carrying, had a significant gradual increase with age. Among males, the increases in scores were significant for the Cooperation/communication and Senses domains and not significant for the Sitting and the Managing domains. For females, there was no clear age gradient for the mental domains, and peaks in reporting difficulties were found in the age groups 44–46 and 84–86.

Education

NFAS scores decreased with more years of education, indicating better self-reported functional ability (Table 5). With the exception of the Senses domain these differences were significant. When splitting the data in males and females and age cohorts, the association between education and Sitting domain disappeared ($p > 0.05$). For the other three physical domains and for the Managing domain, the gradient for education remained evident ($p < 0.05$) in age cohorts 44–76 among males and 24–56 among females. For the Cooperation/communication

Table 3: Descriptive statistics for NFAS items and domain scores by gender (N = 1620)

Domain/item[a]	All (mean)	(SD)	Males (mean)	(SD)	Females (mean)	(SD)	p-value[b]
Walking/standing:	**1.25**	**0.47**	**1.16**	**0.37**	**1.30**	**0.51**	**<0.001**
1. Standing	1.19	0.48	1.12	0.38	1.24	0.54	<0.001
2. Walking less than a kilometre on flat ground	1.19	0.56	1.11	0.43	1.25	0.63	<0.001
3. Walking more than a kilometre on flat ground	1.32	0.74	1.21	0.59	1.41	0.84	<0.001
4. Walking on different surfaces	1.24	0.56	1.16	0.46	1.31	0.62	<0.001
5. Going up and down stairs	1.33	0.63	1.22	0.53	1.41	0.69	<0.001
6. Going shopping for your groceries	1.18	0.49	1.09	0.37	1.25	0.56	<0.001
7. Putting on your shoes and socks	1.21	0.48	1.19	0.43	1.23	0.52	0.33
Holding/picking up things	**1.14**	**0.32**	**1.08**	**0.19**	**1.18**	**0.36**	**<0.001**
8. Picking up a coin from a table with your fingers	1.10	0.37	1.07	0.30	1.13	0.41	<0.001
9. Holding and turning a steering wheel	1.06	0.35	1.01	0.11	1.10	0.46	<0.001
10. Driving a car	1.14	0.57	1.04	0.03	1.23	0.72	<0.001
11. Preparing food	1.10	0.39	1.07	0.37	1.12	0.40	<0.001
12. Writing	1.11	0.38	1.07	0.31	1.15	0.43	<0.001
13. Performing everyday tasks on your own	1.15	0.43	1.07	0.30	1.20	0.50	<0.001
14. Engaging in your leisure activities	1.30	0.65	1.21	0.53	1.36	0.72	<0.001
15. Putting on and taking off your clothes	1.13	0.39	1.09	0.30	1.16	0.44	0.001
Lifting/carrying	**1.23**	**0.46**	**1.12**	**0.30**	**1.31**	**0.52**	**<0.001**
16. Lifting an empty soda bottle crate from the floor	1.15	0.51	1.07	0.32	1.21	0.62	<0.001
17. Carrying shopping bags in your hands	1.23	0.55	1.08	0.33	1.35	0.65	<0.001
18. Carrying a little sack/backpack on your shoulders or back	1.20	0.56	1.09	0.36	1.29	0.67	<0.001
19. Pushing and pulling with your arms	1.31	0.56	1.19	0.47	1.41	0.69	<0.001
20. Cleaning your house	1.33	0.64	1.18	0.50	1.44	0.71	<0.001
21. Washing your clothes	1.16	0.49	1.13	0.48	1.17	0.50	0.02
Sitting	**1.10**	**0.32**	**1.05**	**0.22**	**1.14**	**0.35**	**<0.001**
22. Sitting on a kitchen chair	1.08	0.34	1.05	0.25	1.12	0.39	<0.001
23. Riding as a passenger in a car	1.06	0.27	1.03	0.20	1.08	0.32	<0.001
24. Riding as a passenger on public transport	1.15	0.54	1.07	0.36	1.22	0.64	<0.001
Managing	**1.25**	**0.41**	**1.19**	**0.35**	**1.29**	**0.44**	**<0.001**
25. Staying alert and being able to concentrate	1.26	0.50	1.20	0.46	1.30	0.53	<0.001
26. Working in groups	1.18	0.52	1.14	0.46	1.22	0.57	0.01
27. Guiding others in their activities	1.19	0.56	1.14	0.50	1.23	0.60	0.001
28. Managing everyday responsibility	1.15	0.41	1.10	0.35	1.19	0.46	<0.001
29. Managing everyday stress and strains	1.33	0.58	1.23	0.50	1.40	0.46	<0.001
30. Managing to take criticism	1.34	0.61	1.27	0.56	1.40	0.64	<0.001
31. Managing to control your anger and aggression	1.29	0.53	1.25	0.49	1.32	0.56	0.03
Cooperation/communication	**1.18**	**0.32**	**1.16**	**0.28**	**1.18**	**0.32**	**0.25**
32. Remembering things	1.42	0.61	1.39	0.58	1.45	0.62	0.05
33. Understanding spoken messages	1.21	0.48	1.21	0.46	1.21	0.49	0.81
34. Understanding written messages	1.07	0.31	1.06	0.29	1.08	0.33	0.06
35. Speaking	1.07	0.28	1.05	0.25	1.08	0.31	0.03
36. Participating in a conversation with many people	1.19	0.49	1.19	0.48	1.19	0.50	0.71
37. Using the telephone	1.07	0.32	1.05	0.26	1.09	0.35	0.06
Senses	**1.05**	**0.22**	**1.03**	**0.17**	**1.06**	**0.25**	**0.02**
38. Watching television	1.05	0.24	1.03	0.18	1.06	0.28	0.01
39. Listening to the radio	1.04	0.26	1.03	0.20	1.05	0.30	0.20
Total scores	**1.20**	**0.31**	**1.13**	**0.21**	**1.24**	**0.34**	**<0.001**

[a] Items use a four-point scale of no difficulty, some difficulty, much difficulty and could not do it. Domain scores and total scores are calculated by adding item responses and dividing by the number of items completed.
[b]Mann Whitney U-test.

Table 4: Domain and total NFAS scores in males and females for different age groups (N = 1620)

Domain	Age groups 24–26	34–36	44–46	54–56	64–66	74–76	84–86	p-value[a]
N of females	90	254	158	202	108	73	19	
Walking/standing	1.15	1.21	1.28	1.31	1.46	1.52	2.17	<0.001
Holding/picking up things	1.08	1.13	1.20	1.20	1.22	1.31	1.88	<0.001
Lifting/carrying	1.12	1.22	1.36	1.34	1.40	1.46	2.11	<0.001
Sitting	1.09	1.11	1.15	1.17	1.12	1.13	1.70	<0.001
Managing	1.22	1.28	1.38	1.29	1.29	1.27	1.65	0.01
Cooperation/communication	1.17	1.18	1.24	1.16	1.15	1.19	1.44	0.01
Senses	1.03	1.06	1.10	1.05	1.01	1.01	1.28	0.01
Total scores	1.14	1.19	1.28	1.25	1.28	1.33	1.80	<0.001
N of males	62	175	143	154	109	61	12	
Walking/standing	1.04	1.07	1.16	1.18	1.26	1.32	1.60	<0.001
Holding/picking up things	1.04	1.04	1.06	1.10	1.12	1.12	1.27	<0.001
Lifting/carrying	1.05	1.05	1.10	1.17	1.18	1.18	1.67	<0.001
Sitting	1.03	1.03	1.04	1.09	1.05	1.02	1.08	0.60
Managing	1.15	1.18	1.18	1.23	1.20	1.14	1.52	0.16
Cooperation/communication	1.10	1.12	1.13	1.16	1.24	1.24	1.30	<0.001
Senses	1.01	1.01	1.02	1.04	1.04	1.08	1.08	0.003
Total scores	1.07	1.08	1.12	1.16	1.18	1.18	1.42	<0.001

[a]Kruskal-Wallis test with age group as grouping variable

domain the association with functional ability was significant among females aged 24–46, but not among males.

Discussion

The Norwegian Function Assessment Scale (NFAS) was developed by an expert group to ensure that the instrument has content validity, as a measure of functional ability relevant to the working population. With just 39 items the NFAS is suitable for inclusion in population surveys with minimum respondent burden and take an estimated ten minutes to complete. The instrument seems to be acceptable to the general population in Norway, even though the response rate was relatively modest in some age cohorts. The response rate represents a potential study limitation as we do not know the possible effect imposed by the non-respondents. Compared with national population data [13], the study sample included fewer persons in the youngest and the oldest cohort. Since these two groups are at the opposite ends of the functional ability continuum, the effects on scores might to some extent be cancelled out. Further, more females than males returned the questionnaire, which might have led to poorer scores than if all responded. On the other hand, this effect may have been lessened by the higher percentage of persons with education at university level in the sample compared to the distribution of educational level in the whole population [13].

Table 5: Domain and total NFAS scores in participants with different education levels (N = 1620)

Domain	Years of education ≤9	10 to 12	≥13	p-value[a]
N[b]	299	776	524	
Walking/standing	1.42	1.25	1.14	<0.001
Holding/picking up things	1.24	1.14	1.08	<0.001
Lifting/carrying	1.36	1.25	1.13	<0.001
Sitting	1.14	1.11	1.07	<0.001
Managing	1.38	1.26	1.17	<0.001
Cooperation/communication	1.25	1.19	1.11	<0.001
Senses	1.06	1.05	1.03	0.06
Total scores	1.31	1.20	1.12	<0.001

[a] Kruskal-Wallis test with years of education as grouping variable
[b] Missing data about education for 21 persons

Levels of missing data were within acceptable limits. However, a few items had a high percentage of missing values, which is probably because there was no "not applicable" option. When a participant considered a functional activity irrelevant, he or she would probably have left this item unanswered. Some items could have been irrelevant for the two oldest cohorts since many of these participants have retired from work or do not drive a car. Including a

not applicable option might have lowered missing values for some items.

The NFAS was originally developed for persons of working age. The small number of participants in the two oldest age cohorts, and the poorer data quality among these respondents due to more missing values and irrelevant items imply that caution should be exercised when using these normative data on groups outside the working age. Otherwise, the data quality was acceptable.

Reliability

The level of Cronbach's alpha was acceptable with two of the domains only just failing to meet the criterion of 0.70 for use in groups of people [15]. The participants received the test and the retest questionnaires about two weeks apart. In this way the recall bias might be minimal, but there may have been a real change in health related function. Functional health status is also likely to show some day-to-day variation. For the most part, mean changes were fairly evenly distributed between improvements and deteriorations.

The total proportions of agreement in this test-retest was high compared to a study examining test-retest reliability of COOP/WONCA [17]. Compared to a further test-retest study using the COOP/WONCA charts [18], the weighted kappa values were slightly lower. The ICC values for domains indicated substantial to almost perfect agreement, and all met the reliability standard of 0.70 for use in groups [15]. Compared with other studies using the SF-36 [19,20], ICC values were similar. Overall the test-retest reliability is acceptable.

Normative data

As expected, the data were highly skewed indicating that a large proportion of the population did not experience difficulties with functional activities. One in three respondents reported no difficulty on all items indicating excellent functional ability, and the remaining two thirds reported a variety from minor to major difficulties with different functional activities. The population seems to have most problems with remembering and least problems with their senses. Walking/standing and Managing domain have the highest scores, whereas Senses and Sitting the lowest. The items, watching television and listening to the radio, had very low scores, indicating that very few respondents reported difficulties with this. However, problems with these senses are important aspects in relation to work.

Men reported higher functional ability than women on most items. The findings of previous studies differ somewhat, which may, at least partly, be due to the use of different instruments and the aspects of health that they measure. Of the studies looking at functional health status using the SF-36, five had a similar conclusion [21-26], whereas one study did not [27]. According to one study using the COOP/WONCA charts, males reported better functional ability than females on the first four of the six charts [28]. The report by Grammenos [4], did not show systematically significant differences between the percentages of men and women of working age in the European Union reporting disability.

The significant age gradient in physical domains and the non-gradient in mental domains found in this study follows previous research [21,23,24,26-28]. Grammenos [4] also found a strong non-linear age gradient in the reported disability prevalence rates in the European Union. In our study, females aged 44–46 reported more difficulties on mental domains than younger or older females, the exception being the oldest cohort. For males, a peak at the age group 54–56 was found for the Managing domain only. These findings are supported by the results from a study by Hensing et al [29] showing that the cumulative incidence of sickness absence for a psychiatric diagnosis was highest among those aged 45–59. The association between age and functional ability seems to be more complex in mental domains than in physical domains.

In this study, the length of education was significantly related to functional ability level with better levels among the persons with the highest levels of education. This finding is supported by previous studies [22-25]. In the European Union report [4], education was inversely associated with disability in all countries. Further, positive correlations between income and health, and a presence of collinearity between education, income and socio-economic status were reported. After adjusting for gender and age, we only found associations between educational level and reported functional ability for some subgroups in our study, whereas Sullivan et al [23] reported significant gradients after adjusting for age. The relations between age, gender, education, and income are often difficult to disentangle. In older generations of women, their well-being is more likely to be influenced by their husbands' education and income. The lack of association between functional ability and education in younger men is likely explained by young men's general high functional levels. We propose that a normative population data set must take age, gender, and education into account.

Comparisons with sick-listed persons

Comparing this population study data with data from the sample with 386 Norwegians sick-listed for six weeks [11], the population sample scores are lower than for the sick-listed persons. The largest difference for domains is for the Lifting/carrying domain (1.23 vs. 1.85), and for Walking/

standing, Holding/picking up things and Managing there are 0.31 to 0.35 differences in domain scores for the two samples. Senses had the lowest mean value in both samples (1.05 vs. 1.13). The total score for the sick-listed persons was 1.52 as compared to 1.20 for this population study sample. Looking at single items, the four with the highest difference between the two samples, were engaging in your leisure activities (1.39 vs. 2.28), cleaning your house (1.33 vs. 2.23), carrying shopping bags in your hands (1.23 vs. 1.96) and managing everyday stress and strains (1.33 vs. 1.99). These four functional activities seem to imply much more difficulties for the sample of sick listed than for the normal population. In the sample of 386 sick-listed persons [11] no significant differences between males and females nor any age gradient were found, as opposed to the normal population where females and older persons report more difficulties with functional activities than males and younger persons.

Conclusion

This study presents population scores on the NFAS by gender, age and length of education. Data quality, internal consistency and test-retest reliability were acceptable. The main findings were that females, older persons and persons with lower levels of education reported more functional problems than males, younger persons and persons with higher levels of education. A large proportion of the respondents reported no difficulty for most items and very few answered that they could not do it. The domains, in which the respondents reported most problems with functional activities, were Walking/standing, Lifting/carrying and Managing. These data will serve as basis for the development of national population norms.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

NØ planned and designed the study, performed some of the statistical analysis, drafted the manuscript and coordinated the study. SB planned and designed the study, participated in the interpretation of results and in drafting and revising the manuscript. AG helped in the interpretation of the results and participated in drafting the manuscript. JSB performed most statistical analysis and reviewed the manuscript. BN participated in planning and designing the study, collected the data and participated in drafting the manuscript. PG participated in the planning and design of the study, interpretation of the results and in drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The study is part of The Functional Assessments Project financed by The Ministry of Labour and Social Inclusion. It was carried out in collaboration with The Ullensaker Study 2004 (financed by the University of Oslo and the Trygve Gythfeldt Fund).

References

1. Organisation for Economic Co-operation and Development (OECD): **Transforming Disability into Ability. Policies to Promote Work and Income Security for Disabled People.** Paris, OECD Publishing; 2003.
2. Norges offentlige utredninger: **Sykefravær og uførepensjonering. Et inkluderende arbeidsliv.** *Volume NOU 2000:27*. Oslo, Statens forvaltningstjenester, Seksjon statens trykking; 2000.
3. World Health Organization: **ICF-International Classification of Functioning, Disability, and Health.** Geneva, World Health Organization; 2001.
4. Grammenos S: **Illness, disability and social inclusion.** Dublin: European Foundation for the Improvement of Living and Working Conditions; 2003.
5. Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E: **A quantitative approach to perceived health status: a validation study.** *J Epidemiol Community Health* 1980, **34:**281-286.
6. Ware JE: *SF-36 Health Survey Manual and Interpretation Guide* Boston: The Health Institute New England Medical Center; 1993.
7. Social Security Office: *A guide to incapacity benefit – the personal capability assessment* Leeds: DWP Communications; 2002.
8. Ilmarinen J: **Ageing workers in the European Union – Status and promotion of work ability, employability and employment.** Helsinki: Institute of Occupational Health, Ministry of Social Affairs and Health, Ministry of Labour; 1994.
9. Kessler RC, Stang PE: *Health and Work Productivity. Making the Business Case for Quality Health Care* Chicago and London: The University of Chicago Press; 2006.
10. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, Pronk N, Simon G, Stang P, Ustun TB, Wang P: **The World Health Organization Health and Work Performance Questionnaire (HPQ).** *J Occup Environ Med* 2003, **45:**156-174.
11. Brage S, Fleten N, Knudsrod OG, Reiso H, Ryen A: **[Norwegian Functional Scale--a new instrument in sickness certification and disability assessments].** *Tidsskr Nor Laegeforen* 2004, **124:**2472-2474.
12. Magnussen L: **Returning disability pensioners back to work. Dissertation for PhD-degree.** Department of Education and Health Promotion, University of Bergen; 2007.
13. Statistics Norway: **StatBank Norway.** 2006 [http://www.ssb.no].
14. Fleiss JK: **Statistical Methods for Rates and Proportions, 2nd edn.** New York: Wiley; 1981.
15. Nunnally JC, Bernstein IH: *Psychometric theory* 3rd ed edition. New York: McGraw-Hill; 1994.
16. Landis JR, Koch GG: **The measurement of observer agreement for categorical data.** *Biometrics* 1977, **33:**159-174.
17. Kinnersley P, Peters T, Stott N: **Measuring functional health status in primary care using the COOP-WONCA charts: acceptability, range of scores, construct vaildity, reliability and sensitivity to change.** *Br J Gen Pract* 1994, **44:**545-549.
18. Bentsen BG, Natvig B, Winnem M: **Questions you didn't ask? COOP/WONCA Charts in clinical work and research. World Organization of Colleges, Academies and Academic Associations of General Practitioners/Family Physicists.** *Fam Pract* 1999, **16:**190-195.
19. Hurst NP, Ruta DA, Kind P: **Comparison of the MOS short form-12 (SF12) health status questionnaire with the SF36 in patients with rheumatoid arthritis.** *Br J Rheumatol* 1998, **37:**862-869.
20. Ruta DA, Hurst NP, Kind P, Hunter M, Stubbings A: **Measuring health status in British patients with rheumatoid arthritis: reliability, validity and responsiveness of the short form 36-item health survey (SF-36).** *Br J Rheumatol* 1998, **37:**425-436.
21. Brazier JE, Harper R, Jones NM, O'Cathain A, Thomas KJ, Usherwood T, Westlake L: **Validating the SF-36 health survey questionnaire: new outcome measure for primary care.** *BMJ* 1992, **305:**160-164.
22. Jenkinson C, Coulter A, Wright L: **Short form 36 (SF36) health survey questionnaire: normative data for adults of working age.** *BMJ* 1993, **306:**1437-1440.

23. Sullivan M, Karlsson J: **The Swedish SF-36 Health Survey III. Evaluation of criterion-based validity: results from normative population.** *J Clin Epidemiol* 1998, **51:**1105-1113.
24. Loge JH, Kaasa S: **Short form 36 (SF-36) health survey: normative data from the general Norwegian population.** *Scand J Soc Med* 1998, **26:**250-258.
25. Jenkinson C, Stewart-Brown S, Petersen S, Paice C: **Assessment of the SF-36 version 2 in the United Kingdom.** *J Epidemiol Community Health* 1999, **53:**46-50.
26. Hopman WM, Towheed T, Anastassiades T, Tenenhouse A, Poliquin S, Berger C, Joseph L, Brown JP, Murray TM, Adachi JD, Hanley DA, Papadimitropoulos E: **Canadian normative data for the SF-36 health survey. Canadian Multicentre Osteoporosis Study Research Group.** *CMAJ* 2000, **163:**265-271.
27. Blake C, Codd MB, O'Meara YM: **The Short Form 36 (SF-36) Health Survey: normative data for the Irish population.** *Ir J Med Sci* 2000, **169:**195-200.
28. Bruusgaard D, Nessioy I, Rutle O, Furuseth K, Natvig B: **Measuring functional status in a population survey. The Dartmouth COOP functional health assessment charts/WONCA used in an epidemiological study.** *Fam Pract* 1993, **10:**212-218.
29. Hensing G, Brage S, Nygard JF, Sandanger I, Tellnes G: **Sickness absence with psychiatric disorders--an increased risk for marginalisation among men?** *Soc Psychiatry Psychiatr Epidemiol* 2000, **35:**335-340.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2458/7/278/prepub

II

doi:10.1186/1477-7525-6-14

Health and Quality of Life Outcomes

BioMed Central

Research

Open Access

A randomised comparison of a four- and a five-point scale version of the Norwegian Function Assessment Scale

Nina Østerås*[1], Pål Gulbrandsen[2,3], Andrew Garratt[4], Jūratė Šaltytė Benth[2,3], Fredrik A Dahl[2], Bård Natvig[1] and Søren Brage[1]

Address: [1]Section of Occupational Health and Social Insurance Medicine, Institute of General Practice and Community Health, Faculty of Medicine, University of Oslo, Norway, [2]Helse Øst Health Services Research Centre, Akershus University Hospital, Norway, [3]Faculty of Medicine, University of Oslo, Norway and [4]Institute of Health Management and Health Economics, University of Oslo, Norway

Email: Nina Østerås* - nina.osteras@medisin.uio.no; Pål Gulbrandsen - pal.gulbrandsen@ahus.no; Andrew Garratt - andrew.garratt@kunnskapssenteret.no; Jūratė Šaltytė Benth - jurate@ahus.no; Fredrik A Dahl - fredrik.dahl@ahus.no; Bård Natvig - bard.natvig@medisin.uio.no; Søren Brage - soren.brage@medisin.uio.no

* Corresponding author

Published: 15 February 2008

Health and Quality of Life Outcomes 2008, **6**:14 doi:10.1186/1477-7525-6-14

Received: 4 October 2007
Accepted: 15 February 2008

This article is available from: http://www.hqlo.com/content/6/1/14

Abstract

Background: There is variation in the number of response alternatives used within health-related questionnaires. This study compared a four-and a five-point scale version of the Norwegian Function Assessment Scale (NFAS) by evaluating data quality, internal consistency and validity.

Methods: All inhabitants in seven birth cohorts in the Ullensaker municipality of Norway were approached by means of a postal questionnaire. The NFAS was included as part of The Ullensaker Study 2004. The instrument comprises 39 items derived from the activities/participation component in the International Classification for Functioning, Disabilities and Health (ICF). The sample was computer-randomised to either the four-point or the five-point scale version.

Results: Both versions of the NFAS had acceptable response rates and good data quality and internal consistency. The five-point scale version had better data quality in terms of missing data, end effects at the item and scale level, as well as higher levels of internal consistency. Construct validity was acceptable for both versions, demonstrated by correlations with instruments assessing similar aspects of health and comparisons with groups of individuals known to differ in their functioning according to existing evidence.

Conclusion: Data quality, internal consistency and discriminative validity suggest that the five-point scale version should be used in future applications.

Background

The measurement of functional ability is important in many contexts. While there often seems to be agreement as to the content of instruments for evaluation of function, there is relatively less consensus about the scaling of items. Item scaling vary in the number of response categories, the wording of category options and the use of all-point (where all categories are defined) or end-point (where only end-points are defined) scales [1,2]. The majority of health status and patient-reported outcome measures use all-point defined scales with between two and seven categories, the most popular being five-point

scales including the agree/disagree Likert format. The generic Short Form 36-item (SF-36) Health Survey [3] uses five-point scales for seven of the eight health scales it includes. Other generic instruments such as the Nottingham Health Profile (NHP) [4] and EuroQol EQ-5D [5] use two- and three-point scales respectively. In the WHO Health and Work Performance Questionnaire, functional status is reported using different scales with between four and 11 points [6].

It has been argued that seven-point response scales are the maximum number that individuals are able to process [7] and some authors have advocated their use [8]. However, such scales are not widely used possibly because of the difficulty of finding suitable adjectives when seven all-point defined scales are used. Seven categories are also harder to fit across a page of A4 with a reasonably sized typeface. However, if the number of alternatives is less than the rater's ability to discriminate, the result may be a loss of information [2,9]. There is evidence that the reduction in reliability from ten to seven categories is quite small, but the use of five categories reduces the reliability by about 12 percent [2]. Hence it is argued that the minimum number of categories should be in the region of five to seven [2]. One review concluded that seven plus or minus two appears to be a reasonable range for the optimal number of response alternatives [9]. More recently, it was found that respondents preferences were highest for a ten-point scale followed by seven-point and nine-point scales [10]. The respondents rated scales with five, seven and ten response categories as relatively easy to use. Scales with two, three or four response categories were rated as relatively quick to use, but were unfavourable in terms of the extent to which they allowed the respondents to express their feelings adequately. If a scale does not allow respondents to express themselves, they may become frustrated or demotivated and the quality of their responses may decrease [10].

Previous research has shown that the greater the number of response options, the more reliable the scale is likely to be [11]. Simulations of categorization error have consistently shown that correlation between true values and scale scores increase with the number of response options [12]. Scales with relatively few response alternatives tend to generate scores with comparatively little variance, thereby limiting the magnitude of correlations with other scales [13,14]. The reduction in reliability is most severe for scales with four categories or less, but tends to level off once seven or more options are available. However, there is often a trade-off between scale reliability and ease of administration [11]. One study using the NHP indicated that the psychometric performance and patient acceptability was improved by using a five-point scale instead of the original shorter response format [15].

Following a recent systematic review, it was recommended that future research designs should allocate respondents to different versions of a questionnaire to compare approaches to item scaling [1]. Our study considered two different all-point defined scales using four and five response alternatives. The Norwegian Functional Assessment Scale (NFAS) was included in a large Norwegian population study on musculoskeletal pain, The Ullensaker Study 2004, to obtain self-reported levels of functional ability. Eligible persons were randomised to receive NFAS with the original four-point scale or a five-point scale.

The aim of this study was to compare the original four-point with the new five-point scale version by evaluating validity of the NFAS in a population. This will determine which version should be used in the future applications.

Methods

Study setting and sample

Ullensaker is a rural community which had 23,700 inhabitants in 2004. There are no major differences between the population of Ullensaker and the general population of Norway with respect to demographic characteristics [16]. In 2004, postal questionnaires, which included the NFAS along with questions relating to musculoskeletal pain, were sent to all 6108 inhabitants in Ullensaker municipality in the birth cohorts 1918–20, 1928–30, 1938–40, 1948–50, 1958–60, 1968–70 and 1978–80. Reminders were sent at eight weeks.

The sample was computer-randomised by an external company to either the four-point or the five-point scale version, herein referred to as the NFAS-4 and the NFAS-5. The Ullensaker Study questionnaire also included the Dartmouth COOP Functional Health Assessment Charts/WONCA(COOP/WONCA), General Health Questionnaire-20 (GHQ-20), Standardized Nordic Questionnaire, work ability, sickness absenteeism, and occupation.

The Regional Committee for Medical Research Ethics and The Norwegian Data Inspectorate approved the study.

The Norwegian Function Assessment Scale (NFAS)

The Norwegian Function Assessment Scale (NFAS) is a self-report instrument developed by an expert group in social insurance in 2000 and is designed to assess the need for rehabilitation, adjustment of work demands among sick-listed persons as well as the rights to social security benefits [17]. The scale comprises 39 items derived directly from the activities/participation dimension in the International Classification of Functioning, Disability and Health (ICF) [18]. The items are relevant for assessing physical and mental functioning in working life, some relating to activities of daily living. The NFAS starts with

the question "Have you had difficulty doing the following activities during the last week?" and respondents report 39 activities using a four-point scale: no difficulty, some difficulty, much difficulty, could not do it. The five all-point defined scale was developed to be more congruent with the qualifiers in the activities/participation dimension of ICF [19]: no difficulty, mild difficulty, moderate difficulty, much difficulty and could not do it.

Based on the results of principal component analysis from the previous study with sick-listed persons [17], the items form seven domains: Walking/standing (7 items), Holding/picking up things (8 items), Lifting/carrying (6 items), Sitting (3 items), Managing (7 items), Cooperation/communication (6 items), Senses (2 items). These domains have evidence for validity in sick listed persons [17]. The main application of the NFAS is likely to be social insurance. Hence it was decided to keep the domains from the earlier study with sick-listed persons [17]. It should, however, be anticipated that principal component analysis based on data from the general population in Ullensaker will yield somewhat different results. The first four and the last three domains are intuitively grouped into physical and mental domains respectively. Domain scores are calculated by adding the item scores and dividing by the number of items completed. NFAS total scores are calculated by adding all 39 item scores and dividing by the number of items completed. Low scores indicate good functional ability.

COOP/WONCA

COOP/WONCA [20] is a generic health status measure, where functional status is self-reported with a time frame of the previous two weeks. It comprises six charts: Physical fitness, Feelings, Daily activities, Social activities, Overall health and Change in health. Each chart has five response alternatives with pictorial representations. The present study used an optional Pain chart in place of the Change in health chart.

General Health Questionnaire (GHQ-20)

Psychological distress during the last two weeks was measured by the GHQ-20 [21], a widely used screening instrument for measuring non-psychotic psychiatric illness in a general population. Items are scored as the original GHQ score in a bi-modal fashion (0-0-1-1) [22].

Work ability was assessed by one question "To what degree is your ability to perform your ordinary work reduced today: hardly reduced at all, not much reduced, moderately reduced, much reduced and very much reduced" [23]. Respondents were asked to report whether they had experienced any pain or discomfort in ten different body regions during the previous week [24]. Sickness absenteeism was assessed by asking the respondents if they had been sick-listed during the previous year: no, less than 1 week, between 1–8 weeks, more than 8 weeks. Occupation was assessed with the categories: employed, housekeeping/full-time household work, unemployed, medical rehabilitation, disability pension, retired or student.

Statistical analyses

Data quality

The two versions of the NFAS were compared for levels of missing data, and floor and ceiling effects, which were expressed as percentages.

Tests of scaling assumptions

Internal consistency was assessed by item-total correlation and Cronbach's alpha. Item-total correlation coefficients should meet 0.40 standard. Cronbach's alpha was considered acceptable for group comparisons when the coefficient exceeded 0.70 [25]. Item discriminant validity was assessed by analyzing correlations between the items and their domains (item-total) and between the items and the other domains (item-other) to see if the former was at least two standard errors higher than the latter, thereby indicating definite scaling success [26].

Construct validity

We hypothesised that scores from conceptually related domains of NFAS would correlate higher than scores of unrelated domains. We also hypothesised that NFAS scores would correlate higher with conceptually corresponding aspects of the COOP/WONCA, GHQ and Work Ability than with non-corresponding aspects. Correlation coefficients among measures of the same attribute should fall in the midrange of 0.40 – 0.80 [2].

It was hypothesised that those having a disability pension or rehabilitation benefit due to disease and those reporting being sick-listed previous year, would report lower functional ability. We also compared domain scores between those reporting musculoskeletal pain last week without mental distress (original GHQ score <4) and those with mental distress (original GHQ score ≥ 4) but no musculoskeletal pain. It was hypothesised that females, older persons and persons with shorter education would report lower functional ability than the males, younger persons and persons with longer education. Since data are categorical, non-parametric tests for independent samples were used to compare subgroups.

Results

Sample characteristics

Of the 6108 questionnaires posted, 3325 (54.4%) were returned. The response rate was lower for males ($p < 0.001$) and young or very old persons ($p < 0.001$) (Table 1). The response rates for the two versions were 54.0% for

Table 1: Response rates by age and gender for the NFAS-4 and the NFAS-5 (N = 3325)

	NFAS-4		NFAS-5	
	N (%)	Response rate %	N (%)	Response rate %
Females	905 (55.9)	60.0	919 (53.9)	58.8
Males	715 (44.1)	48.0	786 (46.1)	50.8
All	1620	54.0	1705	54.8
Age:				
24–26	150 (9.3)	33.3	169 (9.9)	37.6
34–36	429 (26.5)	49.9	521 (30.6)	53.7
44–46	301 (18.6)	54.2	301 (17.7)	54.2
54–56	358 (22.1)	68.4	327 (19.2)	62.5
64–66	219 (13.5)	66.2	239 (14.0)	72.2
74–76	132 (8.1)	66.8	120 (7.0)	60.8
84–86	31 (1.9)	37.8	28 (1.6)	34.1

NFAS-4 and 54.8% for NFAS-5. 55 participants in birth cohort 1968–70 randomised to the NFAS-4 were erroneously mailed the NFAS-5 version. Hence, the subsamples differed significantly regarding age ($p < 0.05$), but not on any other background variables. Excluding the birth cohort 1968–1970 did not affect the results.

Data quality

For respondents to the NFAS-4 and NFAS-5, there were no missing data for 78.5% and 82.4% respectively. All items had more missing data for the NFAS-4 than NFAS-5 (Table 2). The mean levels of missing data for individual items in the NFAS-4 and NFAS-5 were 3.3% and 2.6% respectively, which was statistically significant ($p < 0.01$). The same items within both versions had the highest percentage of missing values.

Item responses were skewed towards no difficulty for both versions (Table 2). The percentage of respondents reporting no difficulty for all 39 items was 33.1% in the NFAS-4 and 30.6% in the NFAS-5. In the general the NFAS-4 items had larger floor and ceiling effects than NFAS-5 items; some differences were statistically significant ($p < 0.05$) (Table 2). The third response alternative in NFAS-4 and the fourth in NFAS-5 had exact the same wording, "much difficulty", but the percentage response was lower in NFAS-5 than in NFAS-4 for 24 items.

Scaling assumptions

All items in both versions met the 0.40 criterion for item-total correlation with the exception of the two items in the "senses" domain in NFAS-4 (Table 3). In all domains, item-total correlation coefficients were higher within the NFAS-5 than within NFAS-4, and this difference was significant for 35 items.

All items, except four in the NFAS-4 and one in the NFAS-5, met the item-discriminant validity criterion. Cronbach's alpha for two of the NFAS-4 and one of the NFAS-5 domains just failed to meet the 0.70 criterion (Table 3). Cronbach's alphas were significantly higher for NFAS-5 across the first six domains and the total score.

Construct validity

For both versions, scores from conceptually related domains of NFAS correlated higher than scores of unrelated domains (Table 4). The NFAS-5 produced the largest correlations between domains and between domains and total scores, which was significant ($p < 0.05$) for 15 items and four domains.

NFAS scores correlated higher with conceptually corresponding aspects of the COOP/WONCA, GHQ and Work Ability than with non-corresponding aspects for both versions (Table 4). The Sitting and Senses domains had relatively low correlations with these items or scales. The correlation coefficients were similar for the two versions. With only one exception, all the correlations hypothesized as being high, were over 0.40, indicating that the same construct was being measured by the NFAS and the external standard.

Both versions discriminated between persons anticipated to report different levels of functional ability, including persons with disability pension or medical rehabilitation, persons reporting sickness absence, and persons with physical versus mental symptoms (Table 5).

For both versions, a decline in physical functional ability was significantly associated with increasing age ($p < 0.05$). With one exception, males reported significantly better functional ability ($p < 0.001$) for both versions. With the exception of the Senses domain for the NFAS-4, a significant education gradient was found for both versions ($p < 0.001$).

Table 2: Missing data, means and end effects for NFAS-4 and NFAS-5 items (N = 3325)

		Missing %		Domain/item scores (mean)		Floor %[a]		Ceiling %[a]	
		NFAS-4	NFAS-5	NFAS-4	NFAS-5	NFAS-4	NFAS-5	NFAS-4	NFAS-5
Walking/standing				**1.25**	**1.37**	**61.1**	**62.1**	**0.2**	**0.2**
Standing	1	3.0	2.6	1.19	1.29	84.9	83.2	0.3	0.2
Walking less than a kilometre on flat ground	2	4.6	3.5	1.19	1.30	87.5	84.3**	1.6	1.6
Walking than a kilometre on flat ground	3	3.8	2.8	1.32	1.44	80.6	79.1	4.3	3.2
Walking on different surfaces	4	3.6	3.3	1.24	1.35	81.0	80.1	0.8	0.7
Going up and down stairs	5	2.5	2.1	1.33	1.48	75.0	73.6	1.0	0.3*
Going shopping for your groceries	6	3.2	2.4	1.18	1.30	86.2	82.5**	0.6	1.0
Putting on your shoes and socks	7	1.9	1.8	1.21	1.36	81.6	78.1*	0.3	0.1
Holding/picking up things				**1.14**	**1.23**	**67.5**	**67.5**	**0.1**	**0.1**
Picking up a coin from a table with your fingers	8	2.5	1.9	1.10	1.17	91.6	89.5*	0.1	0.2
Holding and turning a steering wheel	9	5.3	4.9	1.06	1.13	96.3	93.3***	0.9	1.6
Driving a car	10	6.1	4.9	1.14	1.24	93.0	90.3**	3.2	4.1
Preparing food	11	2.5	2.0	1.10	1.16	92.3	89.9*	0.8	0.7
Writing	12	2.2	1.7	1.11	1.18	90.9	88.9	0.2	0.4
Performing everyday tasks on your own	13	2.2	2.3	1.15	1.24	87.9	84.5**	0.4	0.4
Engaging in your leisure activities	14	3.7	3.0	1.30	1.42	78.8	76.7	2.1	1.9
Putting on and taking off your clothes	15	2.2	1.9	1.13	1.20	88.7	86.1*	0.3	0.2
Lifting/carrying				**1.23**	**1.36**	**64.6**	**64.7**	**0.3**	**0.1**
Lifting an empty soda bottle crate from the floor	16	2.6	2.0	1.15	1.23	90.5	87.6**	1.7	1.3
Carrying shopping bags in your hands	17	2.4	1.8	1.23	1.31	82.1	82.1	1.1	0.6
Carrying a little sack/backpack on your shoulders or back	18	2.8	2.3	1.20	1.33	85.8	81.7**	1.8	1.7
Pushing and pulling with your arms	19	3.0	1.9	1.31	1.43	76.0	75.8	1.1	1.1
Cleaning your house	20	3.0	2.1	1.33	1.50	75.2	72.8	1.6	1.6
Washing your clothes	21	3.3	2.9	1.16	1.29	88.6	83.9***	1.3	1.6
Sitting				**1.10**	**1.19**	**87.0**	**82.2**	**0.1**	**0.1**
Sitting on a kitchen chair	22	2.5	1.8	1.08	1.16	93.2	89.7***	0.2	0.2
Riding as a passenger in a car	23	3.5	2.6	1.06	1.12	95.2	91.6***	0.2	0.2
Riding as a passenger on public transport	24	4.5	3.2	1.15	1.25	90.8	86.9**	2.1	1.9
Managing				**1.25**	**1.43**	**53.2**	**46.3**	**0.1**	**0.0**
Staying alert and being able to concentrate	25	2.7	2.2	1.26	1.40	77.3	72.7**	0.2	0.4
Working in groups	26	9.0	6.2	1.18	1.33	86.4	80.6***	1.4	1.3
Guiding others in their activities	27	9.3	7.1	1.19	1.34	86.7	80.6***	2.0	1.8
Managing everyday responsibility	28	3.3	2.9	1.15	1.30	87.6	80.0***	0.2	0.5
Managing everyday stress and strains	29	3.3	2.5	1.33	1.53	72.5	66.1***	0.4	0.7
Managing to take criticism	30	4.3	2.9	1.34	1.54	72.0	63.6***	0.9	0.5
Managing to control your anger and aggression	31	2.2	1.9	1.29	1.49	74.4	65.2***	0.5	0.3
Cooperation/communication				**1.18**	**1.32**	**58.7**	**49.8**	**0.0**	**0.1**
Remembering things	32	2.5	1.9	1.42	1.67	63.5	55.3***	0.5	0.3
Understanding spoken messages	33	2.7	2.1	1.21	1.39	81.6	71.2***	0.3	0.1
Understanding written messages	34	2.5	1.9	1.07	1.16	94.0	88.4***	0.3	0.2
Speaking	35	2.3	1.9	1.07	1.17	93.7	87.6***	0.0	0.1
Participating in a conversation with many people	36	2.6	2.1	1.19	1.35	84.3	77.4***	0.7	0.5
Using the telephone	37	1.9	1.5	1.07	1.15	94.2	90.9***	0.2	0.4
Senses				**1.05**	**1.09**	**94.7**	**91.3**	**0.0**	**0.0**
Watching television	38	2.0	1.6	1.05	1.10	96.1	93.0***	0.0	0.1
Listening to the radio	39	2.0	1.9	1.04	1.09	96.8	94.0***	0.3	0.1
Total score				**1.20**	**1.31**	**33.1**	**30.6**	**0.0**	**0.0**

[a]End effects for the NFAS-4 and NFAS-5 are compared, * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 3: Mean item-total correlation and Cronbach's alpha for domain scores in the NFAS-4 and the NFAS-5 (N = 3325)

	Mean item-total correlation		Cronbach's alpha[a]	
	NFAS-4	NFAS-5	NFAS-4	NFAS-5
Walking/standing	0.74	0.79	0.91	0.93***
Holding/picking	0.55	0.65	0.82	0.88***
Lifting/carrying	0.70	0.77	0.89	0.92***
Sitting	0.53	0.60	0.66	0.74***
Managing	0.66	0.72	0.87	0.91***
Cooperation/communication	0.60	0.66	0.81	0.85***
Senses	0.27	0.53	0.69	0.69
Total scores	0.62	0.70	0.95	0.96**

[a] Cronbach's alpha values for NFAS-4 and NFAS-5 are compared, * p < 0.05; ** p < 0.01; *** p < 0.001

Applying age-stratified analyses, the results for data quality, scaling assumptions and construct validity remained stable.

Discussion

Both versions demonstrated low levels of missing data and skewed response distribution, but the NFAS-4 had more missing values and larger end effects than NFAS-5. The NFAS-5 demonstrated better internal consistency and item-discriminant validity than the NFAS-4, although the results were acceptable for both versions. All a priori hypotheses were met, which strongly supports the construct validity of the scale for both versions. Both versions discriminated similarly well between groups with different levels of health status and between known groups in the population.

Table 4: Correlation[a] between NFAS, COOP/WONCA, GHQ-20 and Work ability for the NFAS-4 and the NFAS-5 (N = 3325)

NFAS-4	**Norwegian Function Assessment Scale**							**COOP/WONCA**			**GHQ-20**	**Work ability**
N = 1620	**Walk./stand.**	**Hold./pick.**	**Lift./carry.**	**Sitting**	**Manag.**	**Coop./ Comm.**	**Senses**	**Phys. fitness**	**Feelings**	**Overall health**		
Walking/standing								**0.46**	0.30	0.58	0.36	0.50
Holding/picking up things	**0.67**							**0.38**	0.32	0.53	0.37	0.52
Lifting/carrying	**0.65**	**0.69**						**0.40**	0.33	0.54	0.39	0.50
Sitting	**0.51**	**0.53**	**0.51**					0.26	0.26	0.40	0.29	0.37
Managing	0.46	0.49	0.49	0.38				0.26	**0.61**	0.58	**0.62**	0.42
Cooperation/ communication	0.37	0.40	0.39	0.26	**0.66**			0.26	**0.42**	0.45	**0.46**	0.34
Senses	0.25	0.26	0.27	0.22	0.24	0.33		0.11	0.16	0.20	0.18	0.20
Total scores	0.77	0.75	0.76	0.52	0.79	0.69	0.29	0.46	0.50	**0.69**	0.56	**0.56**
NFAS-5	**Norwegian Function Assessment Scale**							**COOP/WONCA**			**GHQ-20**	**Work ability**
N = 1705	**Walk./stand.**	**Hold./pick.**	**Lift./carry.**	**Sitting**	**Manag.**	**Coop./ comm.**	**Senses**	**Phys. fitness**	**Feelings**	**Overall health**		
Walking/standing								**0.51**	0.25	0.57	0.36	0.51
Holding/picking up things	**0.73**							**0.41**	0.27	0.54	0.37	0.56
Lifting/carrying	**0.73**	**0.74**						**0.44**	0.28	0.55	0.40	0.58
Sitting	**0.59**	**0.60**	**0.63**					0.34	0.24	0.43	0.32	0.41
Managing	0.51	0.54	0.54	0.48				0.29	**0.56**	0.59	**0.61**	0.46
Cooperation/ communication	0.43	0.47	0.44	0.40	**0.72**			0.28	**0.42**	0.48	**0.47**	0.38
Senses	0.30	0.34	0.32	0.33	0.36	0.42		0.19	0.18	0.27	0.25	0.26
Total scores	0.76	0.76	0.76	0.60	0.83	0.76	0.38	0.45	0.46	**0.67**	0.55	**0.57**

[a]Spearman's correlation
For all correlation coefficients: p < 0.001.
Bold numbers indicate apriori hypothesized associations with high correlation coefficients.

Table 5: Domain scores for different groups of the study population for the NFAS-4 and the NFAS-5 (N = 3325)

	NFAS-4						NFAS-5					
	Disability pension/ rehab.	All others	Sickness absence	No sickness absence	Phys. probl. only	Mental probl. only	Disability pension/ rehab.	All others	Sickness absence	No sickness absence	Phys. probl. only	Mental probl. only
N	196	1414	425	644	603	57	190	1500	461	701	641	76
Walking/ standing	1.66	1.19***	1.22	1.09***	1.20	1.10*	2.13	1.28***	1.34	1.12***	1.33	1.11***
Holding/ picking	1.39	1.11***	1.15	1.04***	1.10	1.05	1.74	1.16***	1.18	1.06***	1.18	1.10**
Lifting/ carrying	1.64	1.18***	1.24	1.09***	1.20	1.06**	2.15	1.26***	1.33	1.11***	1.29	1.12**
Sitting	1.34	1.07***	1.09	1.03***	1.08	1.03	1.64	1.13***	1.16	1.05***	1.14	1.05
Manag.	1.59	1.20***	1.30	1.13***	1.16	1.39***	2.04	1.35***	1.45	1.23***	1.31	1.55*
Coop./ comm.	1.36	1.15***	1.18	1.09***	1.12	1.29***	1.69	1.27***	1.31	1.19***	1.26	1.33
Senses	1.16	1.03***	1.04	1.01***	1.03	1.03	1.24	1.08***	1.09	1.04*	1.07	1.07
Total scores	1.49	1.15***	1.20	1.08***	1.15	1.16	1.91	1.24***	1.30	1.13***	1.25	1.22

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; Mann Whitney U-test

Data quality

The response rates and the low levels of missing data show that both versions of the NFAS are acceptable to the population. A few items had a high percentage of missing values, which is probably because there was no "not applicable" option. Significantly less missing data for the NFAS-5 than the NFAS-4 is some indication that the respondents found it easier choosing a suitable response from the five-point scale. This finding is supported by Nagata et al. [27], who compared feasibility of health measurement response scales using four, five and seven categories and a visual analog scale. The level of missing data was least and the responder preference was highest, for the five-point scale version.

Since the NFAS data are skewed towards higher levels of functioning, the larger end effects for NFAS-4 have to be considered when the instrument is used to discriminate between different levels of functioning or to assess changes in functioning over time. It is likely that NFAS-4 will not be as responsive to changes in functioning, simply because it has fewer response options that individuals can use to indicate that their functioning has changed.

It might be anticipated that the response alternative, "much difficulty", along with the two end categories would show similar percentages in the two versions. This was not found. Hence, the responses did not seem to be affected by the wording or anchoring of the response alternatives.

Internal consistency and validity

The internal consistency values were similar to widely used instruments including the SF-36 [28,29,29-33] and the NHP [15]. Our item-other domain correlation coefficients were comparable with other study results using the SF-36 in a study including rheumatoid arthritis patients [34] and a population study [29].

Regarding construct validity, different time perspectives in the questioning for the different scales could influence possible associations since Work Ability concerns today, NFAS last week, COOP/WONCA and GHQ the last two weeks. However, all a priori hypotheses correlation coefficients met the 0.4 – 0.8 standard. Other studies have obtained similar correlation coefficients between NHP and SF-36 scales [15,34] or between SF-36 scale scores and comparable item or domain scores from other questionnaires [32,35]. Regarding the ability to discriminate between groups with different levels of health status, comparable results were found for the SF-36 [30-33,35]. A gender difference was found in several studies [28,30-32,35-37], but not all [33,38]. The finding of a physical age gradient is supported by several studies [28,32,33,35-38], and an education gradient has also been found in previous research [28,30,31,35,38].

The NFAS-5 demonstrated somewhat higher internal consistency and item-discriminant validity values compared to the NFAS-4. The majority of this difference could probably be attributed to the fact that correlation between true values and scale scores increase with the number of response options [12], but it is not known whether this explains the whole difference in correlation coefficient values.

Future applications of the NFAS

The items in the NFAS are derived directly from the activities/participation dimension in the ICF. The ICF use a five-point scale for their qualifiers and the clinical checklists. This supports the use of the NFAS-5. The NFAS-5 had lower levels of missing data than the NFAS-4 which may indicate higher responder acceptability. The NFAS-5 gen-

erally performed better than the NFAS-4 in relation to the psychometric tests. Therefore the five-point scale is recommended in future applications of the NFAS. The main drawback in changing to a new response format is that it precludes direct comparisons between previous and new research. However, following our study results, we believe that the evidence supports changing the NFAS response format to a five-point scale.

Strengths and limitations

This study' strengths include the randomised design, the large study sample, the good data quality and the thorough testing of validity against other standards. The moderate response rate and that all data is self-reported, represent study limitations. An external, unrelated variable would have strengthened validity assessment. With the present study design it was not possible to ask the respondents about their preferences [10] or to determine the sensitivity to change, the responsiveness of the scale. However, the low mean missing values may indicate acceptability among respondents.

Conclusion

The data quality of NFAS is high with acceptable internal consistency and good construct validity. In choosing between the four-point and the five-point scale, it should be noted that while construct validity and discriminative ability are comparable, both data quality, internal consistency and discriminative validity suggest that the five-point scale is to be preferred in future applications of the NFAS.

Abbreviations

GHQ-20: The General Health Questionnaire-20 items; ICF: The International Classification of Functioning, Disability and Health; NFAS: The Norwegian Function Assessment Scale; SF-36: The generic Short Form 36-item Health Survey

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

NØ planned and designed the study, performed some of the statistical analysis, drafted the manuscript and coordinated the study. PG participated in the planning and design of the study, interpretation of the results and in drafting the manuscript. AG helped in the interpretation of the results and participated in drafting the manuscript. JSB performed most statistical analysis and reviewed the manuscript. FAD assisted statistical analysis and reviewed the manuscript. BN participated in planning and designing the study, collected the data and participated in drafting the manuscript. SB planned and designed the study, participated in the interpretation of results and in drafting and revising the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The study is part of The Functional Assessments Project financed by The Ministry of Labour and Social Inclusion. It was carried out in collaboration with The Ullensaker Study 2004 (financed by the University of Oslo and the Trygve Gythfeldt Fund).

References

1. McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N, Thomas R, Harvey E, Garratt A, Bond J: **Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients.** *Health Technol Assess* 2001, **5:**1-256.
2. Streiner DL, Norman GR: *Health measurement scales a practical guide to their development and use* Third edition. Oxford, Oxford University Press; 2003.
3. Ware JE: *SF-36 Health Survey Manual and Interpretation Guide* Boston, The Health Institute New England Medical Center; 1993.
4. Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E: **A quantitative approach to perceived health status: a validation study.** *J Epidemiol Community Health* 1980, **34:**281-286.
5. Group EQL: **EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group.** *Health Policy* 1990, **16:**199-208.
6. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, Pronk N, Simon G, Stang P, Ustun TB, Wang P: **The World Health Organization Health and Work Performance Questionnaire (HPQ).** *J Occup Environ Med* 2003, **45:**156-174.
7. Miller GA: **The magical number seven plus or minus two: some limits on our capacity for processing information.** *Psychol Rev* 1956, **63:**81-97.
8. Guyatt GH, Townsend M, Berman LB, Keller JL: **A comparison of Likert and visual analogue scales for measuring change in function.** *J Chronic Dis* 1987, **40:**1129-1133.
9. Cox EP: **The Optimal Number of Response Alternatives for a Scale: A Review.** *J Marketing Research* 1980, **17:**407-422.
10. Preston CC, Colman AM: **Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences.** *Acta Psychol (Amst)* 2000, **104:**1-15.
11. Avis NE, Smith KW: *Conceptual and methodological issues in selecting and developing quality of life measures. In: Advances in medical sociology (Fitzpatrick, R, editor).* London, JAI Press Inc.; 2006:255-80.
12. Nishisato S, Torii Y: **Effects of categorizing continuous normal variables on product-moment correlation.** *Japanese Psychological Research* 1970, **13:**45-49.
13. Martin WS: **Effects of Scaling on Correlation Coefficient - Test of Validity.** *Journal of Marketing Research* 1973, **10:**316-318.
14. Chang L: **A Psychometric Evaluation of 4-Point and 6-Point Likert-Type Scales in Relation to Reliability and Validity.** *Applied Psychological Measurement* 1994, **18:**205-215.
15. Cleopas A, Kolly V, Perneger TV: **Longer response scales improved the acceptability and performance of the Nottingham Health Profile.** *J Clin Epidemiol* 2006, **59:**1183-1190.
16. StatisticsNorway: **StatBank Norway.** 2006 [http://www.ssb.no].
17. Brage S, Fleten N, Knudsrod OG, Reiso H, Ryen A: **[Norwegian Functional Scale--a new instrument in sickness certification and disability assessments].** *Tidsskr Nor Laegeforen* 2004, **124:**2472-2474.
18. World Health Organization: **ICF-International Classification of Functioning, Disability, and Health.** Geneva, World Health Organization; 2001.
19. World Health Organization: **ICF Checlist. Version 2.1a, Clinical Form for International Classification of Functioning, Disability and Health.** 2007 [http://www.who.int/classifications/icf/site/checklist/icf-checklist.pdf]. Accessed 15-11-2007
20. Nelson E, Wasson J, Kirk J, Keller A, Clark D, Dietrich A, Stewart A, Zubkoff M: **Assessment of function in routine clinical practice: description of the COOP Chart method and preliminary findings.** *J Chronic Dis* 1987, **40 Suppl 1:**55S-69S.

21. Goldberg DP: *Manual of the General Health Questionnaire* Edited by: NFER-Nelson . Windsor; 1978.
22. McDowell I: *Measuring Health. A Guide to Rating Scales and Questionnaires* Third edition. Oxford, University Press; 2006.
23. Reiso H, Nygard JF, Brage S, Gulbrandsen P, Tellnes G: **Work ability assessed by patients and their GPs in new episodes of sickness certification.** *Fam Pract* 2000, **17(2):**139-144.
24. Kuorinka I, Jonsson B, Kilbom A, Vinterberg H, Biering-Sorensen F, Andersson G, Jorgensen K: **Standardised Nordic questionnaires for the analysis of musculoskeletal symptoms.** *Appl Ergon* 1987, **18:**233-237.
25. Nunnally JC, Bernstein IH: *Psychometric theory* 3rd ed edition. New York, McGraw-Hill; 1994.
26. Kaasa S, Bjordal K, Aaronson N, Moum T, Wist E, Hagen S, Kvikstad A: **The EORTC core quality of life questionnaire (QLQ-C30): validity and reliability when analysed with patients treated with palliative radiotherapy.** *Eur J Cancer* 1995, **31A:**2260-2263.
27. Nagata C, Ido M, Shimizu H, Misao A, Matsuura H: **Choice of response scale for health measurement: comparison of 4, 5, and 7-point scales and visual analog scale.** *J Epidemiol* 1996, **6:**192-197.
28. Loge JH, Kaasa S: **Short form 36 (SF-36) health survey: normative data from the general Norwegian population.** *Scand J Soc Med* 1998, **26:**250-258.
29. Sullivan M, Karlsson J, Ware JE Jr.: **The Swedish SF-36 Health Survey--I. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden.** *Soc Sci Med* 1995, **41:**1349-1358.
30. Jenkinson C, Coulter A, Wright L: **Short form 36 (SF36) health survey questionnaire: normative data for adults of working age.** *BMJ* 1993, **306:**1437-1440.
31. Jenkinson C, Stewart-Brown S, Petersen S, Paice C: **Assessment of the SF-36 version 2 in the United Kingdom.** *J Epidemiol Community Health* 1999, **53:**46-50.
32. Brazier JE, Harper R, Jones NM, O'Cathain A, Thomas KJ, Usherwood T, Westlake L: **Validating the SF-36 health survey questionnaire: new outcome measure for primary care.** *BMJ* 1992, **305:**160-164.
33. Blake C, Codd MB, O'Meara YM: **The Short Form 36 (SF-36) Health Survey: normative data for the Irish population.** *Ir J Med Sci* 2000, **169:**195-200.
34. Loge JH, Kaasa S, Hjermstad MJ, Kvien TK: **Translation and performance of the Norwegian SF-36 Health Survey in patients with rheumatoid arthritis. I. Data quality, scaling assumptions, reliability, and construct validity.** *J Clin Epidemiol* 1998, **51:**1069-1076.
35. Sullivan M, Karlsson J: **The Swedish SF-36 Health Survey III. Evaluation of criterion-based validity: results from normative population.** *J Clin Epidemiol* 1998, **51:**1105-1113.
36. Hopman WM, Towheed T, Anastassiades T, Tenenhouse A, Poliquin S, Berger C, Joseph L, Brown JP, Murray TM, Adachi JD, Hanley DA, Papadimitropoulos E: **Canadian normative data for the SF-36 health survey. Canadian Multicentre Osteoporosis Study Research Group.** *CMAJ* 2000, **163:**265-271.
37. Bruusgaard D, Nessioy I, Rutle O, Furuseth K, Natvig B: **Measuring functional status in a population survey. The Dartmouth COOP functional health assessment charts/WONCA used in an epidemiological study.** *Fam Pract* 1993, **10:**212-218.
38. Grammenos S: **Illness, disability and social inclusion.** Dublin, European Foundation for the Improvement of Living and Working Conditions; 2003.

III

Implementing structured functional assessments in general practice for persons with long-term sick leave: a cluster randomised controlled trial. Nina Østerås, Pål Gulbrandsen, Jūratė Šaltytė Benth, Dag Hofoss and Søren Brage. BMC Family Practice 2009, 10:31

doi:10.1186/1471-2296-10-31

The electronic version of this article is the complete one and can be found online at:
http://www.biomedcentral.com/1471-2296/10/31

Implementing structured functional assessments in general practice for persons with long-term sick leave: a cluster randomised controlled trial

Nina Østerås[1*], Pål Gulbrandsen[2,3], Jūratė Šaltytė Benth[2,3], Dag Hofoss[2], Søren Brage[1]

[1] Section of Occupational Health and Social Insurance Medicine, Institute of General Practice and Community Health, Faculty of Medicine, University of Oslo, Norway

[2] HØKH, Research Centre, Akershus University Hospital, Norway

[3] Faculty Division Akershus University Hospital, University of Oslo, Norway

[*]Corresponding author

Email addresses:

NØ: nina.osteras@medisin.uio.no

PG: pal.gulbrandsen@ahus.no,

JSB: jurate@ahus.no,

DH: dag.hofoss@ahus.no

SB: soren.brage@medisin.uio.no

Abstract

Background: The increasing attention on functional assessments in medical and vocational rehabilitation requires a focus change for the general practitioners (GP) into paying attention to patient resources, possibilities and coping instead of symptoms, problems and limitations. The GPs report difficulties in performing the requested explicit functional assessments. The purpose of this study was to implement a structured method in general practice for assessing functional ability in persons with long-term sick leave. The study aim was to evaluate intervention effects on important GP parameters; knowledge, attitudes, self-efficacy towards functional assessments and knowledge about patient work factors.

Methods: Fifty-seven GPs were randomly assigned to an intervention or a control group. The intervention group GPs attended an introductory one-day work-shop and implemented structured functional assessments during an eight months intervention period. GP knowledge, GP attitudes, and GP self-efficacy towards functional assessments, as well as GP knowledge of patient work factors, were collected before, after and six months after the intervention period started. Evaluation score-sheets were filled in by both the intervention GPs and their patients immediately after the consultation to evaluate the GPs' knowledge of patient work factors.

Results: The intervention GPs reported increased knowledge (B: 0.56, 95% CI (0.19, 0.91)) and self-efficacy (B: 0.90, 95% CI (0.53, 1.26)) towards functional assessments, and increased knowledge about their patients' workplace (B: 0.75, 95% CI (0.35, 1.15)) and perceived stressors (B: 0.55, 95% CI (0.23, 0.88)) with lasting effects at the second follow-up. No intervention effect was seen in relation to GP attitudes. Both before and after the intervention, the GPs were most informed about physical stressors, and less about mental and work organisational stressors (Guttman's reproducibility coefficient: 0.95 and 1.00). After the consultation, both the intervention GPs and their patients reported that the GPs' knowledge about patient work factors had increased (GP B: 0.60 (95% CI: 0.42, 0.78); patient B: 0.50 (95% CI: 0.34, 0.66)).

Conclusions: Introducing and implementing structured functional assessments in general practice made the GPs capable to assess functional ability of their patients in a structured manner. Intervention effects of increased GP knowledge and GP self-efficacy persisted at the second follow-up.

Background

Assessments of patients' functional ability are necessary in medical and vocational rehabilitation. To an increasing extent, general practitioners (GP) in the European countries are being asked to assess function, in addition to disease and illness, in social security claims [1][2]. This focus on functional ability is unfamiliar to GPs [3]. It represents a shift in their attention from patient symptoms, problems and limitations into resources, possibilities and coping. Earlier, functional assessments have been an implicit part of their practice, whereas at present an explicit communication of functional abilities is required. The GPs reported difficulties and were reluctant to meet this request [3]. This was due to lack of training and guidelines, as well as confusing terminology and insufficient knowledge of specific occupational demands [3]. In 2003 only 35% of the GPs in Norway met the request for functional assessments in sickness certification forms [4]. Additionally, the GPs' procedures for functional assessments are usually non-standardised and strongly influenced by their personal and professional interest in functional assessments and working life in general [3].

Methods for structured functional assessment have been developed and tested in some countries, including England and Finland [5-7], but to our knowledge there is no previous randomised controlled study directed at functional assessments of persons with long-term sick leave in general practice. The many randomised, controlled studies addressing professional educational or quality assurance interventions carried out to improve quality of care, show that active multifaceted approaches are more likely to be effective compared to passive single interventions [8].

Based on these experiences, a structured method for functional assessments of persons with long-term sick leave in general practice was developed and tested by GPs in a cluster randomised controlled trial. The purpose was to provide a tailor-made, structured functional assessment method for GPs in busy and ordinary primary care practices. The method was designed to be appropriate for assessing and communicating functional ability information along with suggestions for workplace adjustments to local social security officers and employers. Intervention effects on patient sick leave will be reported elsewhere.

Study objectives

The first aim of this study was to assess intervention effects on GP knowledge, GP attitudes and GP self-efficacy towards functional assessments. The second aim was to assess intervention effects on GP knowledge about their patients' perceived physical, mental and organisational stressors at the workplace. The third aim related to the patient level and was to assess whether the intervention GPs and their patients had similar evaluations of the GPs' knowledge about patient work factors immediately after the consultation.

Methods

Study setting and sample

With the assistance of the Section of General Practice, University of Oslo, and of local medical consultants, 360 GPs in the south-eastern part of Norway were identified and written invitations were sent in November 2004. The responders were randomly assigned to the intervention or the control group according to a computer generated randomisation list made by an independent researcher. The researchers were not blinded to group allocation.

The intervention GPs were requested to apply the intervention on ten consecutive sick-listed persons. The criteria for including a sick-listed person were: being part-time or full-time sick-listed for between eight and 26 weeks and having good prospects of a return to work, meaning that the GPs should exclude persons they thought were candidates for permanent disability benefits.

Informed written consents were received from all GPs. For reasons of anonymity, no written consent was collected from their patients, but the GPs asked their patients for a verbal informed consent. The Regional Committee for Medical Research Ethics and The Norwegian Data Inspectorate approved the study.

Sample size

Using a table for sample size determination [9] we specified a power of 80% to detect a medium-sized difference of 1.2 standardised effect size in relation to knowledge about functional assessments at the GP level with a significance level of 5%. We found the required sample size to be 22 GPs in each group.

The intervention

The target for the multifaceted intervention was the intervention group GPs. The structured functional assessment method was introduced at a one-day workshop including teamwork and role-playing. The need to practice the assessments as part of the process of the trial was acknowledged, and the project group provided phone support when needed. The workshop was accredited by the Norwegian Medical Association for continuing medical education points. The intervention GPs were requested to apply the intervention method on their patients according to the inclusion and exclusion criteria mentioned above.

The included patients were asked to self-report their functional abilities prior to the GP-consultation using the Norwegian Function Assessment Scale (see Additional file 1). This instrument was developed by an expert group in social insurance in 2000 to assess the need for rehabilitation, adjustment of work demands among sick-listed persons as well as the rights to social security benefits [10]. It comprises 39 items derived from the activities/participation component in the International Classification for Functioning, Disabilities and Health [11]. The items are relevant for assessing physical and mental functioning in working life, some relating to activities of daily living [12,13]. The sick-listed persons were also asked to self-report work exposures and perceived stressors at work prior to the GP-consultation using the Work Description Form (see Additional file 2).

During the consultation, the GP independently assessed the patient's functional abilities on the basis of the two forms, the patient's medical history, clinical findings, and motivation. The assessment was formalised as the Function Assessment Report (see Additional file 3), which was sent to the employer and the local social security office. This whole procedure was expected to take about 40 minutes. The GPs in the control group were requested to assess functional ability as usual during the intervention period: March - October, 2005.

Outcome measures

GP knowledge, attitude, self-efficacy and GP knowledge of patient perceived stressors

The quality of the educational and implementation components of the intervention was measured in the main questionnaire, which was tailor-made for this study by the project group. The main questionnaire included 19 items (see Additional file 4), and the first item mapped self-reported GP knowledge about functional assessments. Four items were constructed to cover GP attitudes towards functional assessments (items no. 4 and 7-10), and

three items were made to assess GP self-efficacy towards performing such functional assessments (items no. 11, 13 and 14). Of the remaining 11 items, five were related to GP knowledge about patient work factors: the workplace (item no. 15), the work tasks (item no. 16), and the perceived stressors at work (items no. 17-19). The last five items were included for other purposes and not relevant in this study (items no. 2, 3, 5, 6, and 12). Knowledge and attitude were assessed since they represent the first and second process stage for adopting new ideas or changing behaviour [14]. Self-efficacy (mastering beliefs) was measured since it might influence the initial decision to perform behaviours, the effort expended, and the length of time the individual will persist in the face of obstacles and aversive experiences [15]. All items in the main questionnaire were scored along a five all-point defined scale from very poor to very good, or from totally disagree to totally agree. There was an inverse scoring for one item (no. 10). A small group of GPs pilot tested the questionnaire beforehand.

The main questionnaire was completed by the randomised GPs at three time points: immediately before (T0) and after the intervention period (T1), and at the follow-up six months later (T2). Two written and one oral reminders were given to non-respondents.

GP and patient evaluations

The evaluation score-sheet was used to measure the performance of the functional assessment method itself and how it influenced the GPs' knowledge about patient work factors (see Additional files 5 and 6). It was filled in by both the GP and the patient immediately after the GP-consultation. The GP and the patient rated the GP's knowledge level on two items using a five all-point defined scale from no knowledge to exceptionally good knowledge. The first and the second item were related to the GP knowledge level before and after the consultation, respectively. The patients' evaluations can be seen as a validation of the GPs' evaluations.

Descriptive data

Information on characteristics of the GPs was collected to allow comparisons with national data: gender, age, speciality in family medicine, working hours per week, number of consultations per day, and list size. For reasons of anonymity, only the consultation date, gender and age for the included patients were registered.

Statistical analyses

The scores from all participating GPs were included in the baseline and the longitudinal analyses, although three, and then four GPs did not return the two follow-up questionnaires, respectively (Figure 1). Two methods of imputing missing values, last-observation-carried-forward (LOCF) and the median imputation, were applied for single items missing and for total questionnaire missing at T1 and T2 (those lost to the follow-ups). This did not change the conclusions, so results from the original non-imputed dataset are presented. Non-parametric and parametric tests for independent samples were used to compare subgroups and to compare participants' descriptive data with national data.

GP knowledge, attitude and self-efficacy towards functional assessments

Confirmatory factor analyses using AMOS [16] were used to test the main questionnaire data against hypothesised model structures. As a result, 11 of the included 14 items sum to form three domains (χ^2 (df=59)=71.645, p=0.125): GP attitudes (items no. 4 and 7-10), GP self-efficacy (items no. 11, 13 and 14) and GP knowledge about patient perceived stressors at work (items no. 17-19). Domain scores for these three domains were calculated by adding the item scores and dividing by the number of items completed. The remaining three items (no. 1, 15 and 16) were used as single items in the following analyses.

Non-parametric tests for two related samples were used to analyse domain and item score changes in attitude, self-efficacy and knowledge between two time points, whereas the linear mixed model for repeated measurements using SPSS (version 14.0.2) was estimated to assess longitudinal score changes. This linear mixed bi-level model was chosen for the longitudinal analyses because it allows missing item values. All variables were treated as fixed effects including the intercept and an interaction variable: time by group. The three domain scores and the three single item scores were used as dependent variables. The covariance (among repeated measures of dependent variables on the same individuals) model was chosen using Akaike's Information Criterion [17], and compound symmetry structure was the best. With compound symmetry covariance it is assumed that the variance is constant across occasions [18], and a close examination of dependent variable correlations between different time points showed low variations. The intervention effect was assessed by the interaction term to analyse if the scores of the two groups differed in time from T0 to T1 (the type III Wald tests, $p< 0.05$). To analyse the stability of the intervention effect at the second follow-up, T2 was compared with T1. All estimates in the multivariate models were adjusted for GP gender, age and number of daily GP consultations.

GP knowledge of patient perceived stressors

To assess the potential grading of GP knowledge about their patients' perceived physical, mental and work organisational stressors at the workplace at T0 and T1, the Guttman's reproducibility coefficient [19] was calculated. Guttman's reproducibility coefficient shows which fraction of the responses to a set of questions designed to measure one dimension that fits the cumulative pattern. It can be read as the chance to predict correctly the responder's answer to any given question on the basis of his/her sum-score (i.e., sum of endorsed items in a set of questions). A Guttman's reproducibility coefficient of 1 means that all responders with a sum-score of 1 achieved their one point on the "easiest" question to agree to, all those who scored 2 points got their points by agreeing to the two "easiest" questions etc. To conclude that the observed data fit a Guttman-scale, the reproducibility coefficient should exceed 0.90 [20].

GP and patient evaluations

Evaluation score-sheet data was analysed by two separate linear mixed models for repeated measurements with patients (level 1) nested within the intervention group GPs (level 2). All variables were treated as fixed effects including the intercept, and compound symmetry structure was used as covariance model because before and after scores were correlated. The dependent variables were the GP-evaluated and the patient-evaluated knowledge scores after the consultation. Estimates were adjusted for the GP-evaluated and patient-evaluated knowledge scores before the consultation, GP and patient gender and age as well as the number of daily GP consultations.

Results

Sample characteristics

Of the 360 GPs invited, 57 (15.8%) agreed to participate (Figure 1). No information was obtained about non-respondents. Missing item values were few, ranging 0.0 – 1.9% for the main questionnaire and 0.8% for the two evaluation score-sheet items. The GPs in the intervention group applied the intervention on a total of 133 sick-listed persons (2-10 per GP). For these patients, the mean age was 44.8 years and the percentage of males was 31.5%. The mean age and the percentage of males among long-term sick-listed persons on a national basis for the same period, was 42.0 years and 37.5% respectively [21]. A small proportion of the intervention GPs needed some phone support and guidance in the beginning.

There was no significant difference between the intervention group and the control group at T0 with respect to background information (Table 1). Compared to all GPs in Norway, the proportion of female GPs and the GPs' mean age in the study sample were slightly higher, but the difference was not significant. The proportion of specialists in family medicine and the list size for the participating GPs were significantly higher than the corresponding national numbers ($p<0.05$)[22].

GP knowledge, attitude and self-efficacy towards functional assessments

No significant difference between the two groups was found at baseline in relation to GP knowledge, GP attitudes, or GP self-efficacy. At T1 the intervention group reported significantly ($p< 0.05$) more knowledge about functional assessments, the patients' workplace, work tasks, and perceived stressors as well as higher self-efficacy regarding functional assessments (Table 2). The change in mean scores ranged 0.5 – 0.8 for the five-point scale. There were ignorable or no changes in mean GP attitude scores. In the control group there were no changes in the scores. The stability of the mean scores was tested at T2, and no significant changes in mean scores were seen in the intervention group, whereas in the control group the mean score for knowledge about perceived stressors increased significantly.

Both crude and adjusted longitudinal analyses were done, but since estimates and standard errors were very similar, only the adjusted estimates are shown in Table 3. A significant ($p< 0.05$) intervention effect was found for GP knowledge about functional assessments, GP self-efficacy, and GP knowledge about the patients' workplace, work tasks, and perceived stressors, but not for GP attitudes. Adjusted estimates ranged 0.6-0.9 for the five-point scale. Increasing GP age was significantly associated with increasing knowledge about the patients' workplace and perceived stressors. The intervention effect remained evident at the second follow-up, except for knowledge about perceived stressors. Seven of the intervention GPs stated at the second follow-up that they had continued to use the structured functional assessment method, although they were no longer equivalently paid for prolonged consultations.

GP knowledge of patient perceived stressors

The GPs reported that they were most informed about their patients' perceived physical stressors at work, less informed about mental and even less informed about work

organisational stressors at work (Guttman's reproducibility coefficient: 0.95) at T0. For the intervention group at T1, the Guttman's reproducibility coefficient was: 1.00.

GP and patient evaluations

Data from 130 pairs of evaluation score-sheets (two pairs were not returned and one patient score-sheet had missing items), 130 filled in by intervention GPs and 130 by their respective patients, was available for statistical analyses. Both the GP and the patients evaluated the GP's knowledge about patient work factors as significantly higher immediately after the consultation, adjusted estimates: 0.57 (95% CI: 0.46, 0.69) and 0.48 (95% CI: 0.38, 0.57), respectively for the five-point scale.

Discussion

Summary of main findings

The use of a structured method for functional assessment in general practice led to significantly increased GP knowledge and higher self-efficacy towards functional assessments. In addition, the GPs showed increased knowledge about the patients' workplace and perceived stressors. The intervention effects sustained at the second follow-up six months later. The GPs were better informed about their patients' physical than about their mental and work organisational perceived stressors. Both the intervention GPs and their patients reported increased GP knowledge about patient work factors as a result of the consultation.

GP knowledge, attitude and self-efficacy towards functional assessments

Earlier studies have shown that active interventions can increase knowledge levels [23-27], although there are exceptions [28]. An increase in self-efficacy has also been reported by others [23,25,26,28]. We found no intervention effect on attitudes towards functional assessments, and failure in changing attitude levels has also been reported by others [27,28]. This could be due to regression to the mean, the phenomenon whereby respondents with extreme values will, for purely statistical reasons, probably give less extreme measurements on other occasions. An alternative explanation is that this study sample of volunteers represented a selected group that already was very positive towards functional ability. Thus, it might have been difficult to achieve further positive changes in attitude, which has also been suggested by others [23,27].

The relatively large increases in GP knowledge and self-efficacy mean scores, 0.6-0.9 on a five-point scale, not only represent statistically significant changes, but probably also reflect clinically relevant changes. The highest change in mean score, 0.9, was found for self-efficacy regarding functional assessments. This increase could be attributed to a combination of increased knowledge about functional assessments along with practice and experience in doing such assessments. Since the intervention provided a "tool" for performing functional assessments, this might have increased the GPs' mastering beliefs.

GP knowledge of patient perceived stressors

An increase in knowledge about the individual patients' workplace and perceived stressors was expected since the functional assessment method requires that the GP collect more information and spend more time on work related issues in the prolonged consultation. The reason for the significant change in mean score for knowledge about perceived stressors in the control group from T1 to T2 is unknown. At the same time, however, there was a non-significant decrease in mean score in the intervention group. These changes in opposite directions may cause the significant estimate for the interaction term in the longitudinal analyses comparing T1 and T2.

The large potential for increasing GP knowledge about the patient perceived work organisational stressors during the functional assessment was not utilized. This could indicate that the GPs feel more competent to handle and address physical, rather than mental and work organisational factors, in their work with persons on long-term sick leave. The structured functional assessment method treats the three as equally important factors, but maybe more focus should have been given to assess work organisational stressors.

GP and patient evaluations

The GPs reported a slightly higher ($p>0.05$) increase than their patients did, in GP knowledge about patient work factors. A previous study has found high agreement between the GP and the patient when assessing work ability [29].

Implications for future research or clinical practice

Most intervention GPs implemented the method on a small number of patients. Possibly, the effects of the intervention could have been greater if the number of patients was higher. The GPs selected the patients themselves according to the inclusion and exclusion criteria.

However, the patient inclusion criteria used in this study, ‘having good prospects of a return to work’, is not very specific. It often relies on a subjective judgement by the GPs, but in our opinion the present method is not appropriate in cases where the patient applies for permanent social benefits. We believe, from our contact with the GPs, that the patient selection was random, but it cannot be excluded that the GPs chose patients known to display compliance. This might have given the GPs a skewed impression of the intervention’s usefulness. However, the patients were representative for long-term sick-listed persons in relation to age and gender.

This method for functional assessment is quite time-consuming compared to a normal GP consultation in busy and ordinary practice, which is estimated to last for 10-20 minutes. The low implementation rate among the GPs indicates that this method is unlikely to be, and should not be, implemented routinely. In our opinion the method should rather be applied selectively with the most relevant application being cases of complex long-term sick leave where the GP recognise the need for a more thorough assessment of the patient. Also, implementation of the method may be initiated by the local social security officer requesting information on functional ability. By providing such information along with suggestions for workplace adjustments, The Function Assessment Report may facilitate an early return to work.

The findings suggest that a one-day workshop, with some phone support, is sufficient to provide the GPs with adequate background information to apply the structured functional assessment method to persons with long-term sick leave. For future work, it would be interesting to have a critical look at what the GPs wrote in the Function Assessment Forms, whether they pointed out patient resources and if they provided suggestions for workplace adjustments to facilitate a quick return to work. Explorations of the patient self-reported functional ability level and work demands in relation to register based sick leave also represent an interesting possibility for future work.

Strengths and limitations

The randomised design minimizes the effect of biases that we were unable to control for, and low levels of missing item values contributed to good data quality in this study. The second follow-up gave us a possibility to assess the stability of the intervention effect, and the

patients' evaluation of the GPs' knowledge level represents a validation of the GPs' own evaluation.

The number of GPs included in the study was low compared to the number of invited GPs, but according to power analyses, the number of GPs in each group was satisfactory. Like in other studies [23,28], it proved difficult to recruit GPs on a voluntary basis for a clinical study.

The sample is representative for general practice with regard to age and gender. At the same time they probably belong to a highly selected group that is more interested in functional assessments than many other GPs are. Such self-selection bias was unavoidable and probably reflects that mainly interested GPs voluntarily seek special skills and utilise structured methods in this domain.

The five intervention GPs that withdrew after the randomisation raises the possibility of post-randomisation selection bias, thus representing a study weakness. As we have no collected data for these persons, we cannot do drop-out analysis for these five GPs. Further, it means that no true intention–to-treat principle can be followed in this study.

The use of self-reporting rather than objective measures for the study outcomes represents another limitation in this study. Along with the lack of blinding of the GPs, it might have led to bias for the positive results in this study. However, it represented a feasible way of measuring different components of the intervention, and objective measures of register based patient sick leave will be reported elsewhere.

Conclusions

This study showed that a structured functional assessment method enhanced the GPs' knowledge about functional assessments and patient work factors, as well as their self-efficacy towards performing functional assessments. A one-day workshop and phone support provided the GPs with adequate background information to apply these assessments to persons with long-term sick leave. The intervention effects sustained at the follow-up six months later.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NØ planned and designed the study, collected the data, performed most statistical analyses, drafted the manuscript and coordinated the study. PG participated in the planning and designing of the study, interpretation of the results and in drafting the manuscript. JSB assisted the statistical analyses and reviewed the manuscript. DH performed the factor analysis, the Guttman's reproducibility coefficient analysis, assisted other statistical analyses and reviewed the manuscript. SB planned and designed the study, participated in the interpretation of results and in drafting and revising the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

We thank Magne Thoresen for statistical advices and Therese Nordberg Hanevold for accurate work in the data collection period. Additionally, we thank all the study participants. The study is part of The Functional Assessments Project financed by The Ministry of Labour and Social Inclusion.

References

1. Organisation for Economic Co-operation and Development (OECD): *Transforming Disability into Ability. Policies to Promote Work and Income Security for Disabled People.* Paris: OECD Publishing; 2003.

2. Brage S, Donceel P, Falez F: **Development of ICF core set for disability evaluation in social security.** *Disabil Rehabil* 2008, **30:** 1392-1396.

3. Krohne K, Brage S: **New rules meet established sickness certification practice. A focus group study on the introduction of functional assessments in Norwegian primary care.** *Scand J Prim Health Care* 2007, **25:** 172-177.

4. Andersen EJ: **[Functional assessments causing trouble? (in Norwegian].** *Tidsskr Nor Laegeforen* 2003, **123:** 2928.

5. Swales K, Craig P: *Evaluation of the Incapacity Benefit Medical Test: In-house report 26.* London: Social Research Branch, Analytical Services Division, Department of Social Security; 1997.

6. Tuomi K, Ilmarinen J, Jahkola A, Katajarinne L, Tulkki A: *Work Ability Index*, 2nd revised edn. Helsinki: Finnish Institute of Occupational Health; 1998.

7. Nygard CH, Eskelinen L, Suvanto S, Tuomi K, Ilmarinen J: **Associations between functional capacity and work ability among elderly municipal employees.** *Scand J Work Environ Health* 1991, **17 Suppl 1:** 122-127.

8. Grimshaw JM, Shirran L, Thomas R, Mowatt G, Fraser C, Bero L *et al*.: **Changing provider behavior: an overview of systematic reviews of interventions.** *Med Care* 2001, **39:** II2-45.

9. Altman DG: *Practical statistics for medical research*. London: Chapman and Hall; 1991.

10. Brage S, Fleten N, Knudsrod OG, Reiso H, Ryen A: **[Norwegian Functional Scale--a new instrument in sickness certification and disability assessments (in Norwegian)].** *Tidsskr Nor Laegeforen* 2004, **124:** 2472-2474.

11. World Health Organization: *ICF-International Classification of Functioning, Disability, and Health*. Geneva: World Health Organization; 2001.

12. Osteras N, Brage S, Garratt A, Benth JS, Natvig B, Gulbrandsen P: **Functional ability in a population: normative survey data and reliability for the ICF based Norwegian Function Assessment Scale.** *BMC Public Health* 2007, **7:** 278.

13. Osteras N, Gulbrandsen P, Garratt A, Benth JS, Dahl FA, Natvig B *et al*.: **A randomised comparison of a four- and a five-point scale version of the Norwegian Function Assessment Scale.** *Health Qual Life Outcomes* 2008, **6:** 14.

14. Rogers EM: *Diffusion of innovation*, 4th edn. New York: The free press; 1995.

15. Bandura A: **Self-efficacy: toward a unifying theory of behavioral change.** *Psychol Rev* 1977, **84:** 117-125.

16. Arbuckle J, Wothke W: *Amos 4.0 User's guide*. Chicago: SmallWaters Corporation; 1999.

17. Akaike H: **A new look at the statistical model identification.** *IEEE Transactions on Automatic Control* 1974, **19:** 716-723.

18. Fitzmaurice GM, Laird NM, Ware JH: *Applied longitudinal analysis*. Hoboken, N.J: Wiley-Interscience; 2004.

19. Massof RW: **Likert and Guttman scaling of visual function rating scale questionnaires.** *Ophthalmic Epidemiol* 2004, **11:** 381-399.

20. Hays RD, Ellickson PL: **Guttman scale analysis of longitudinal data: a methodology and drug use applications.** *Int J Addict* 1990, **25:** 1341-1352.

21. Norwegian Insurance Administration: *[Yearbook of National Insurance Statistics 2003 (in Norwegian)]*. Oslo: NIA; 2003.

22. Norwegian Labour and Welfare Administration. [Principal number report, per 4. quarter 2005 (in Norwegian]. [Online] 10-2-2006 [cited 21-12-2006]: Available from: URL: http://www.nav.no/binary/1073748592/file

23. Sanci LA, Coffey CM, Veit FC, Carr-Gregg M, Patton GC, Day N *et al.*: **Evaluation of the effectiveness of an educational intervention for general practitioners in adolescent health care: randomised controlled trial.** *BMJ* 2000, **320:** 224-230.

24. Forsetlund L, Bradley P, Forsen L, Nordheim L, Jamtvedt G, Bjorndal A: **Randomised controlled trial of a theoretically grounded tailored intervention to diffuse evidence-based public health practice [ISRCTN23257060].** *BMC Med Educ* 2003, **3:** 2.

25. Margalit AP, Glick SM, Benbassat J, Cohen A, Katz M: **Promoting a biopsychosocial orientation in family practice: effect of two teaching programs on the knowledge and attitudes of practising primary care physicians.** *Med Teach* 2005, **27:** 613-618.

26. Short LM, Surprenant ZJ, Harris JM, Jr.: **A community-based trial of an online intimate partner violence CME program.** *Am J Prev Med* 2006, **30:** 181-185.

27. Shuval K, Berkovits E, Netzer D, Hekselman I, Linn S, Brezis M *et al.*: **Evaluating the impact of an evidence-based medicine educational intervention on primary care doctors' attitudes, knowledge and clinical behaviour: a controlled trial and before and after study.** *J Eval Clin Pract* 2007, **13:** 581-598.

28. King M, Davidson O, Taylor F, Haines A, Sharp D, Turner R: **Effectiveness of teaching general practitioners skills in brief cognitive behaviour therapy to treat patients with depression: randomised controlled trial.** *BMJ* 2002, **324:** 947-950.

29. Reiso H, Nygard JF, Brage S, Gulbrandsen P, Tellnes G: **Work ability assessed by patients and their GPs in new episodes of sickness certification.** *Fam Pract* 2000, **17:** 139-144.

FIGURE 1. Flow chart of participants through trial

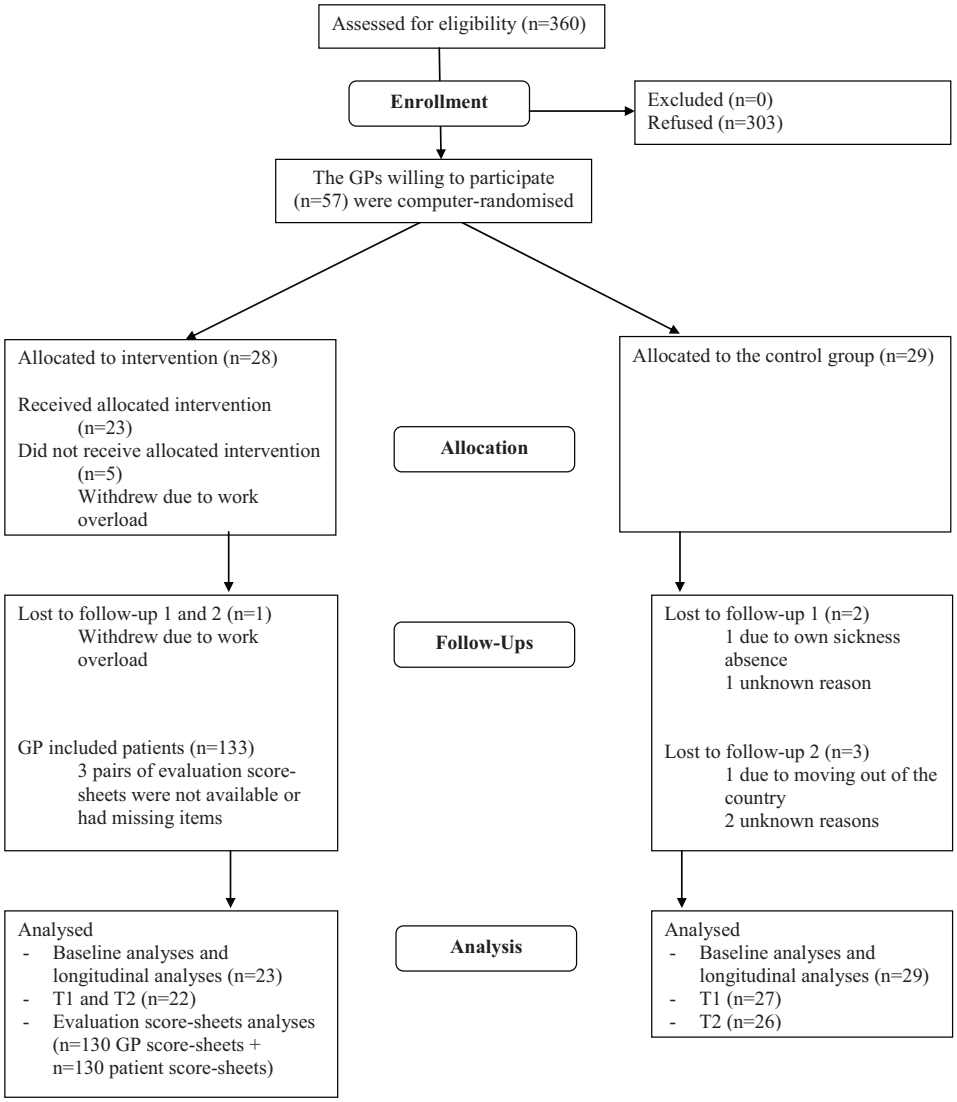


TABLE 1. Sample characteristics of the participating GPs and corresponding national data for general practice GPs, 2005.

	Intervention group (n=23)	Control group (n=29)	Norwegian general practice GPs[1] (n=3757)
Females, n (%)	8 (34.8)	11 (38.0)	1145 (30.5)
Males, n (%)	15 (65.2)	18 (62.1)	2612 (69.5)
Speciality in Family Medicine, n (%)	16 (69.6)	24 (82.8)	2217 (59.0)*
Mean age, y (SD)	49.3 (10.4)	49.5 (8.7)	47.9
Mean Weekly working hours, h (SD)	37.5 (7.2)	41.3 (8.5)	-
Mean daily consultations, n (SD)	21.8 (4.7)	21.0 (4.8)	-
Mean list size, n (SD)	1254.1 (397.4)	1309.8 (210.0)	1189.0*

[1] Numbers from The Norwegian Labour and Welfare Administration [22]

* $p < 0.05$

TABLE 2. Cluster level analyses (GP level) on knowledge, attitudes and self-efficacy in intervention and control groups. Mean scores with 95% confidence intervals at three time points.

		T0	T1	T2
		n=23[a]+29[b]	n=22[a]+27[b]	n=22[a]+26[b]
		Mean (95% CI)	Mean (95% CI)	Mean (95% CI)
GP knowledge about functional assessments	**IG**	**3.1 (3.0, 3.2)**	**3.8 (3.5, 4.1)***	**3.8 (3.5, 4.1)**
	CG	3.2 (3.0, 3.4)	3.3 (3.1, 3.6)	3.4 (3.1, 3.7)
GP attitude towards functional assessments	**IG**	**4.0 (3.7, 4.3)**	**4.1 (3.8, 4.4)**	**4.0 (3.7, 4.3)**
	CG	4.1 (3.9, 4.3)	4.1 (3.8, 4.4)	3.8 (3.5, 4.1)
GP self-efficacy	**IG**	**3.1 (2.8, 3.4)**	**3.9 (3.6, 4.2)****	**3.8 (3.5, 4.1)**
	CG	3.5 (3.3, 3.7)	3.4 (3.1, 3.7)	3.5 (3.2, 3.8)
GP knowledge about the workplace	**IG**	**2.8 (2.5, 3.1)**	**3.6 (3.3, 3.9)***	**3.6 (3.3, 3.9)**
	CG	2.9 (2.6, 3.2)	2.9 (2.6, 3.2)	3.1 (2.8, 3.4)
GP knowledge about the work tasks	**IG**	**3.1 (2.8, 3.4)**	**3.6 (3.3, 3.9)***	**3.6 (3.4, 3.8)**
	CG	3.1 (2.9, 3.3)	3.1 (2.9, 3.3)	3.4 (3.1, 3.7)
GP knowledge about perceived stressors	**IG**	**3.0 (2.8, 3.2)**	**3.5 (3.3, 3.7)***	**3.3 (3.0, 3.6)**
	CG	2.9 (2.7, 3.1)	2.9 (2.7, 3.1)	3.1 (2.9, 3.3)*

* p< 0.05, ** p< 0.01

The p-values are based on non-parametric tests for related samples between T0 and T1, and T1 and T2.

[a:] Intervention group, [b:] Control group

T0: Immediately before the intervention period started, T1: Immediately after the intervention period ended, T2: Six months after the intervention period ended

IG: Intervention group, CG: Control group

TABLE 3. Cluster level (GP level; n=52) longitudinal analyses on knowledge, attitudes and self-efficacy. Estimates [a] are adjusted for GP age, gender and number of daily consultations.

Dependent and independent variables	Multivariate (adjusted)			
	Estimate	SE	p-value	95% CI
GP knowledge about functional assessments				
Time x group (T0 – T1)	0.56	0.18	0.003	0.19, 0.91
Time x group (T1 - T2)	-0.12	0.18	0.500	-0.48, 0.24
GP gender	-0.05	0.15	0.757	-0.35, 0.26
GP age	0.01	0.01	0.148	-0.01, 0.03
No. of daily consultations	0.02	0.02	0.126	-0.01, 0.06
GP attitude towards functional assessments				
Time x group (T0 – T1)	0.20	0.16	0.231	-0.13, 0.52
Time x group (T1 - T2)	0.13	0.16	0.420	-0.19, 0.46
GP gender	0.17	0.18	0.361	-0.20, 0.54
GP age	-0.00	0.01	0.741	-0.02, 0.02
No. of daily consultations	0.02	0.02	0.402	-0.02, 0.06
GP self-efficacy				
Time x group (T0 – T1)	0.90	0.18	<0.001	0.53, 1.26
Time x group (T1 - T2)	-0.13	0.19	0.478	-0.50, 0.24
GP gender	-0.28	0.19	0.143	-0.66, 0.10
GP age	0.01	0.01	0.419	-0.01, 0.03
No. of daily consultations	0.02	0.02	0.326	-0.02, 0.06
GP knowledge about the workplace				
Time x group (T0 – T1)	0.75	0.20	<0.001	0.35, 1.15
Time x group (T1 - T2)	-0.22	0.20	0.282	-0.63, 0.19
GP gender	-0.02	0.16	0.881	-0.33, 0.28
GP age	0.02	0.01	0.010	0.01, 0.04
No. of daily consultations	0.01	0.02	0.659	-0.02, 0.04
GP knowledge about the work tasks				
Time x group (T0 – T1)	0.39	0.19	0.049	0.02, 0.77
Time x group (T1 - T2)	-0.18	0.20	0.360	-0.57, 0.21
GP gender	0.04	0.14	0.771	-0.24, 0.33
GP age	0.01	0.01	0.091	-0.00, 0.03
No. of daily consultations	0.01	0.01	0.988	-0.03, 0.03
GP knowledge about perceived stressors				
Time x group (T0 – T1)	0.55	0.17	0.001	0.23, 0.88
Time x group (T1 - T2)	-0.41	0.17	0.018	-0.74, -0.07
GP gender	0.13	0.12	0.279	-0.11, 0.37
GP age	0.01	0.01	0.025	0.00, 0.03
No. of daily consultations	0.00	0.01	0.829	-0.02, 0.03

[a] The outcome measures were analysed by linear mixed models for repeated measurements using T1, the control group and male GP scores as reference values.

Additional data files

Additional file 1: PDF-file showing The Norwegian Function Assessment Form

Additional file 2: PDF-file showing The Work Description Form

Additional file 3: PDF-file showing The Function Assessment Report

Additional file 4: PDF-file showing the Main Questionnaire

Additional file 5: PDF-file showing the GP Evaluation Score-sheet

Additional file 6: PDF-file showing the Patient Evaluation Score-sheet

IV

DOI: 10.1177/1403494809357096

Structured functional assessments in general practice increased the use of part-time sick leave: a cluster randomised controlled trial

Nina Østerås[1], Pål Gulbrandsen[2,3], Inger Cathrine Kann[2], Søren Brage[1]

[1] Section for Social Medicine, Institute of General Practice and Community Health, Faculty of Medicine, University of Oslo, Norway;

[2] HØKH, Research Centre, Akershus University Hospital, Norway

[3] Faculty Division Akershus University Hospital, University of Oslo, Norway

Address for correspondence and reprint requests:

Nina Østerås

Section for Social Medicine,

Institute of General Practice and Community Health,

Faculty of Medicine, University of Oslo,

P.O Box 1130 Blindern, N-0318 Oslo, Norway

Phone number: +47 22 85 06 29

Fax number: +47 22 85 06 10

E-mail: nina.osteras@medisin.uio.no

Abstract

Aim: A method for structured functional assessments of persons with long-term sick leave was implemented in a cluster randomised controlled trial in general practice. The aim was to analyse intervention effects on GP sick-listing practice and patient sick leave.

Methods: 57 GPs were randomly assigned to an intervention or a control group. The intervention group GPs learned the method at a one-day workshop including teamwork and role-playing. The control group GPs were requested to assess functional ability as usual during the eight months intervention period in 2005. Outcome measures included duration of patient sick leave episodes, GP prescription of part-time sick leave, active sick leave, and vocational rehabilitation. This data was extracted from a national register.

Results: The GPs in the intervention group prescribed part-time sick leave more often ($p<0.01$) and active sick leave less often ($p=0.04$) than the control group GPs during the intervention period. There was no intervention effect on duration of patient sick leave episodes or on GP prescription of vocational rehabilitation.

Conclusion: Implementing structured functional assessments in general practice made the GPs capable to assess functional ability of persons with long-term sick leave in line with new requests. The intervention GPs' sick-listing practice was changed as they prescribed significantly more part-time and less active sick leave compared to the control group GPs. As a result, more intervention GP patients returned part-time to work compared to control GP patients. No intervention effect was seen on duration of patient sick leave episodes or on prescription of vocational rehabilitation.

Keywords: functional assessment, randomised controlled trial, family practice, sick leave

Word count: 2994

Background

General practitioners (GP) in European countries are to an increasing extent being asked to assess function, in addition to disease and illness, in social security claims (1). Here, in medical and vocational rehabilitation, functional assessments represent a balancing of individual functional abilities against occupational demands and restrictions as a prerequisite for successful reintegration into working life. By assessing functional abilities, the patient's resources, possibilities and coping are emphasized, which may facilitate and encourage an early return to work. Public authorities and insurance companies hope this focus on functional abilities will reduce the number of people claiming benefits and the duration of sick leave.

Functional assessments have always been an implicit part of the GP practice, whereas at present an explicit communication of functional abilities is required. However, the GPs report difficulties with this and are reluctant to meet the request due to lack of training and guidelines as well as confusing terminology and insufficient knowledge of specific occupational demands (2). This might explain the lack of functional ability information in many sickness certification forms (3).

Structured functional assessment methods have been introduced, and are used extensively in some countries, including England and Finland (4-6), but to our knowledge there is no previous randomised controlled trial (RCT) directed at functional assessments in general practice. Several reviews have concluded that interactive multifaceted interventions, preferably including an opportunity to practice skills and the use of follow-up sessions, appear to be effective in changing GP performance (7-9). As new knowledge disperses relatively quickly in the medical societies, the challenge is getting the GPs to translate their knowledge into action and behavioural changes (10).

Based on these experiences, a structured method for GP functional assessments of persons with long-term sick leave was developed. The purpose was to provide a tailor-made, structured method for GPs in busy and ordinary primary care practices for functional assessments and communication of functional ability information to the local social security officer and employers. A cluster RCT was conducted to test the method and assess effects on self-reported GP parameters, GP sick-listing practice, and patient sick leave. The intervention effects on self-reported GP parameters have previously been reported (11).

Aim

The aim of this study was to assess intervention effects on GP sick-listing practice and patient sick leave.

Methods

Study setting and sample

With the assistance of Section of General Practice, University of Oslo, and local medical consultants, 360 GPs in the south-eastern part of Norway were identified and written invitations were sent in November 2004. The responders were randomly assigned to the intervention or the control group according to a computer generated randomisation list made by an independent researcher. The researchers were not blinded to group allocation.

The Regional Committee for Medical Research Ethics and The Norwegian Data Inspectorate approved the study.

Sample size

The sample size for this cluster RCT was determined using an outcome variable from a previous study (11). Using a table for sample size determination (12) we specified a power of 80% to detect a medium-sized difference of 1.2 standardised effect size in relation to knowledge about functional assessments at a significance level of 5%. We found the required sample size to be 22 GPs in each group.

The intervention

The intervention GPs learned the intervention method, structured functional assessments, at a one-day workshop. This workshop included teamwork and role playing and was accredited by the Norwegian Medical Association for continuing medical education points. The intervention GPs were requested to apply the intervention method on ten consecutive persons with long-term sick leave. The criteria for including a sick-listed person were: being part-time or full-time sick-listed for between eight and 26 weeks (57 – 182 days) and having good prospects of a return to work. This means that the GPs should exclude persons they thought were candidates for permanent disability benefits. The control group GPs were requested to assess functional ability as usual during the intervention period from March 1st to October 31st in 2005.

The intervention method included two patient questionnaires, a GP consultation, and the end-product, the Function Assessment Report. Before the consultation, the sick-listed patients self-reported physical and mental functional abilities on the 39 items in the Norwegian Function Assessment Scale (13-15) and work exposures and perceived stressors at work on the Work Description Form. During the consultation, the GP independently assessed the patient's functional abilities on basis of the two questionnaires, the patient's medical history, clinical findings, and motivation. The assessment was formalised as the Function Assessment Report, and the whole procedure took about 40 minutes. Data from these assessments will not be used in this study.

Outcome measurements

We have defined sick leave as medically certified absence from work due to disease, illness or injury. By long-term sick leave, we mean sick leave episodes lasting for more than eight weeks. There were four outcome measures of the GPs' sick-listing practice in this study:

The first: the duration of patient sick leave episodes was defined as the number of calendar days from the first day of sick leave until reported off the sick list. This outcome variable was continuous with range 57 - 365 days (eight weeks - maximum sick leave).

The second: part-time sick leave is a GP certification option allowing the employee to be absent from work for a specified proportion of the working hours or work week. This variable was coded as a binary response variable, as whether the GP prescribed part-time sick leave or not during the sick leave episode.

The third: active sick leave is a GP certification option that enables people on sick leave to attend work doing other tasks than they normally do. The National Insurance Administration provides 100% remuneration of normal wages during the active sick leave period of maximum eight weeks, which is the same compensation level as for ordinary sick leave. The fourth: vocational rehabilitation represents support or allowance that can be granted to persons on sick leave who need to change job or job training because of ill health. Active sick leave and vocational rehabilitation were coded as number of calendar days to GP prescription, measured from when the sick leave episode started. These were continuous variables with range 57 - 365 days.

Information on GP characteristics: gender, age, speciality in Family Medicine, working hours per week, number of consultations per day, and number on the patient list, was collected by a self-administered questionnaire.

The data file

Data on patient sick leave was extracted from the register of The Norwegian Labour and Welfare Administration. For reasons of anonymity, we could not identify the individual sick-listed persons included by the GPs. Therefore, all sick leave episodes for the participating GPs were extracted from the register. Of these sick leave episodes, only episodes reaching duration between eight and 26 weeks in the intervention period from March 1st to October 31st in 2005 were included in the analysis file. For historical reference data, sick leave episodes reaching duration between eight and 26 weeks from March 1st to October 31st in 2004 were also included in the analysis file. The sick leave episodes were followed until the person was reported off the sick list or until the maximum date for receiving sick leave benefits was reached, which is after 365 days. After exclusions (n=712, see Figure 1), the data file contained 4562 sick leave episodes.

The register included ICPC codes for diagnoses, which were recoded into severe (n=124) versus less severe disease. The severe diseases included malignant neoplasm, cardiac failure, severe head injuries, mental retardation and psychoses.

Statistical analyses

Outcome measurement analyses with two-level models were conducted using Stata (version 10.0). Cox proportional hazards survival analysis with standard errors adjusted for GP clusters was used to analyse the duration of patient sick leave episodes and GP prescription of active sick leave and vocational rehabilitation. The patients reported off the sick list before reaching their maximum date and the patients prescribed to active sick leave or vocational rehabilitation, were coded as complete and the others as censored. Part-time sick leave was analysed by a binary response two-level regression model with 4562 sick leave episodes (level 1) nested within the 52 GPs (level 2). All estimates were adjusted for GP and patient gender and age, as well as being classified with a severe disease. Only one significant interaction term was found (Table II).

Results

Of the 360 GPs invited, 57 agreed to participate (Figure 1). No information was obtained about non-respondents. There were no significant differences between the intervention group and the control group at T0 with respect to background information (Table I). Compared to all GPs in Norway, the proportion of female GPs and the GPs' mean age in the study sample were slightly higher, but the difference was not significant. The proportion of specialists in family medicine and the list size for the participating GPs were significantly higher than the corresponding national numbers ($p<0.05$) (16). The GPs in the intervention group applied the intervention method on a total of 133 sick-listed patients (2-10 per GP), and the total number of sick leave episodes included in analyses was 4562. The proportion of female patients and the patients' mean age were similar to national data for long-term sick-listed persons in the same period (Table 1) (17).

Duration of patient sick leave episodes

The mean number of patient sick leave days was similar for the two groups (n=190-191), which indicate no intervention effect on the duration of patient sick leave episodes in this study (Table II). For both groups the mean duration was reduced by five days from the reference period to the intervention period.

Part-time sick leave

The intervention GPs prescribed part-time sick leave significantly ($p<0.05$) more often (OR 1.3) than the control GPs (Table III). The proportion of part-time sick leave increased significantly ($p<0.001$) for both groups from the reference to the intervention period, 48.1% to 63.0% and 47.5% to 56.0% for the intervention and the control group, respectively.

Active sick leave

Significantly ($p<0.05$) less active sick leaves were prescribed by the intervention GPs compared to the control GPs (Table IV). For both groups, the proportion was reduced from the reference to the intervention period, 8.7% to 4.6% and 9.8% to 7.0% for the intervention and the control group, respectively.

Vocational rehabilitation

Vocational rehabilitation was initiated in only a small number of sick leave episodes in both groups, and there was no significant difference between the two groups (HR: 1.04, 95% CI

(0.63, 1.70)). There was a small reduction in percentage prescription from the reference to the intervention period, 4.2% to 3.5% and 3.4% to 3.3% for the intervention and the control group, respectively.

Discussion

Summary of main findings

Implementing structured functional assessments in a cluster RCT in general practice led to changes in the GPs' sick-listing practice. As a result of the intervention, the intervention group GPs prescribed significantly more part-time and less active sick leave compared to the control group GPs. There was no intervention effect on the duration of patient sick leave episodes or on GP prescription of vocational rehabilitation.

Implications of the findings

Sick leave is affected by factors on different structural levels, it has a multifactor origin, and it involves various actors with diverse roles and incentives, which leads to challenges in interpreting the results of sick leave intervention studies (18). The potential success of this study intervention depended on several factors, and among these, changing the GP behaviour was important. However, previous research have shown inconsistent success in changing the GP behaviour (7-9). In our study, the GP behaviour was changed in some, but not all respects, similar to what was reported in a RCT with structured GP assessments of patients with long-term mental illness (19). In contrast to this, an earlier Norwegian study promoting GP prescription of active sick leave for patients with low back pain, showed that the GPs were insignificantly susceptible to the intervention (20).

Previous GP intervention research on duration of sick leave is limited (18), and the results show no conclusive evidence. In studies on patients with mental or emotional distress, one intervention study with occupational physicians showed a shortening of sick leave duration (21), while one general practice study showed no effect on duration (22). Similarly, a cluster RCT in general practice to improve quality of care for patients with low back pain showed no effect on duration of sick leave (23). In our study, however, both the low sample size and the low implementation rate might have been too small to detect intervention effects on duration. Additionally, the low implementation rate might have made the duration of sick leave episodes sensitive to chance variations in the patients' illness, thus minimizing the

intervention effect. The significant interaction term indicates that the intervention effect on duration of sick leave was higher for the female than for the male intervention GPs.

It has been argued that functional assessments may facilitate an early return to the workplace (24), which is in line with our results. However, if the return to work is only part-time, there might be a tendency towards longer duration of sick leave episodes, as reported in one Swedish study (25), but in our study no such prolongation was found. As long as sick leave episodes are not extended, an increased use of part-time sick leave means less total sick leave, cost reduction, and increased production for the society.

The proportion of GP initiated active sick leave in our study naturally decreased as the prescription of part-time sick leave increased. The proportion of initiated vocational rehabilitation was very small for both groups and not affected by the intervention.

New rules for sickness certifications, emphasizing the importance of doing work-related activities during sick leave, were implemented by July 1, 2004, which was towards the end of the reference period. All GPs were also requested to use part-time sick leave more often, and to consider part-time sick leave before active sick leave. Due to the new rules, Norwegian GPs' sick-listing practice changed considerably with fewer sick leave cases, shorter duration of sick leave episodes, increased prescription of part-time sick leave, and decreased prescription of active sick leave (26). The prescription of vocational rehabilitation was reasonably stable (27). All these national trends were reflected in our study, but the intervention effect on part-time and active sick leave came in addition.

Implications for future research or clinical practice

There is an impending risk that persons on long-term sick leave will get increasingly distanced from the work sphere of society, and that return to work may grow difficult as the duration of sick leave episodes increases (28). However, this effect may be counteracted by prescribing part-time sick leave, which leads to sustained or re-established contact with the workplace. According to a recent study, 92% of the persons on part-time sick leave were satisfied with the arrangement and 62% of those on full time sick leave thought that part-time sick leave would be a good thing for themselves (29).

This method for functional assessment is more time-consuming than normal GP consultations, and the low implementation rate among the GPs indicates that it should not be implemented routinely or it should be simplified or shortened. In our opinion the present method is not appropriate in cases where the patient applies for permanent social benefits, whereas in cases of complex long-term sick leave where the GPs recognise the need for a more thorough assessment of the patient, an application is more relevant. For future use, an assessment may be initiated by the local social security officer requesting information on functional ability. By providing such information along with suggestions for workplace adjustments, The Function Assessment Report may facilitate an early return to work.

Strengths and limitations

The main strengths of this study are the randomised design and that the outcome data is not self-reported, but extracted from a national register. We also had the opportunity to collect data from a historical reference period. Hence, we could compare the two GP groups' sick-listing practice at two time points and separate the effects of the legislation change from the intervention effects.

Like in other studies (19;22;23), it proved difficult to recruit GPs on a voluntary basis for a clinical study, and only a moderate proportion of invited GPs chose to participate. However, the sample is representative for general practice with regard to age and gender. At the same time they probably belong to a highly selected group that is more interested in functional assessments than other GPs are. Such self-selection bias was unavoidable and probably reflects that mainly interested GPs voluntarily seek special skills and utilise structured methods in this domain. The randomised design minimizes the effect of other biases that we were unable to control for.

The five intervention GPs that withdrew after the randomisation raises the possibility of post-randomisation selection bias, thus representing a study weakness. As we have no data for these five GPs, we can neither do a drop out analysis, nor follow a true intention–to-treat principle in relation to the GPs.

In this study patient sick leave was used as an indirect measure of the GPs' sick-listing practice, which is accounted for by the application of two-level data models in the statistical analyses. Since we were not allowed to identify the 133 patients that were assessed by the

intervention GPs, we included all sick leave episodes for the participating GPs in the analyses. This might have weakened the intervention results, but in our opinion it may represent a more valid picture of how the GPs' sick-listing practice was changed.

Conclusion

The results of this cluster RCT indicate that implementing a method for structured functional assessments affected the GP sick-listing practice. A one-day workshop provided the GPs with adequate background information to apply structured functional assessments on persons with long-term sick leave in line with new requests. As a result, the intervention GPs prescribed significantly more part-time and less active sick leave, but no intervention effect was seen on duration of patient sick leave episodes or on GP prescription of vocational rehabilitation.

Acknowledgements

We thank Magne Thoresen for statistical advices and Therese Nordberg Hanevold for accurate work in the data collection period. Additionally, we thank all the study participants. The study is part of The Functional Assessments Project financed by The Ministry of Labour and Social Inclusion.

References

(1) Organisation for Economic Co-operation and Development (OECD). Transforming Disability into Ability. Policies to Promote Work and Income Security for Disabled People. Paris: OECD Publishing; 2003.

(2) Krohne K, Brage S. New rules meet established sickness certification practice. A focus group study on the introduction of functional assessments in Norwegian primary care. Scand J Prim Health Care 2007;25(3):172-7.

(3) Andersen EJ. [Functional assessments causing trouble? (In Norwegian]. Tidsskr Nor Laegeforen 2003;123(20):2928.

(4) Swales K, Craig P. Evaluation of the Incapacity Benefit Medical Test: In-house report 26. London: Social Research Branch, Analytical Services Division, Department of Social Security; 1997.

(5) Tuomi K, Ilmarinen J, Jahkola A, Katajarinne L, Tulkki A. Work Ability Index. 2nd revised ed. Helsinki: Finnish Institute of Occupational Health; 1998.

(6) Nygard CH, Eskelinen L, Suvanto S, Tuomi K, Ilmarinen J. Associations between functional capacity and work ability among elderly municipal employees. Scand J Work Environ Health 1991;17 Suppl 1:122-7.

(7) O'Brien MA, Freemantle N, Oxman AD, Wolf F, Davis DA, Herrin J. Continuing education meetings and workshops: effects on professional practice and health care outcomes. Cochrane Database Syst Rev 2001;(2):CD003030.

(8) Grimshaw JM, Shirran L, Thomas R, Mowatt G, Fraser C, Bero L, et al. Changing provider behavior: an overview of systematic reviews of interventions. Med Care 2001 Aug;39(8 Suppl 2):II2-45.

(9) Bero LA, Grilli R, Grimshaw JM, Harvey E, Oxman AD, Thomson MA. Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. The Cochrane Effective Practice and Organization of Care Review Group. BMJ 1998 Aug 15;317(7156):465-8.

(10) Hjortdahl P. [Why and how should the clinical behavior of physicians be changed? (In Norwegian)]. Tidsskr Nor Laegeforen 1994 Aug 30;114(20):2372-3.

(11) Osteras N, Gulbrandsen P, Benth JS, Hofoss D, Brage S. Implementing structured functional in general practice for persons with long-term sick leave: a cluster randomised controlled trial. Accepted BMC Family Practice 2008.

(12) Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.

(13) Brage S, Fleten N, Knudsrod OG, Reiso H, Ryen A. [Norwegian Functional Scale--a new instrument in sickness certification and disability assessments (In Norwegian)]. Tidsskr Nor Laegeforen 2004 Oct 7;124(19):2472-4.

(14) Osteras N, Brage S, Garratt A, Benth JS, Natvig B, Gulbrandsen P. Functional ability in a population: normative survey data and reliability for the ICF based Norwegian Function Assessment Scale. BMC Public Health 2007;7:278.

(15) Osteras N, Gulbrandsen P, Garratt A, Benth JS, Dahl FA, Natvig B, et al. A randomised comparison of a four- and a five-point scale version of the Norwegian Function Assessment Scale. Health Qual Life Outcomes 2008 Feb 15;6(1):14.

(16) Norwegian Labour and Welfare Administration. [Principal number report, per 4. quarter 2005 (In Norwegian]. [Online] 2006 February 10 [cited 2006 Dec 21];Available from: URL: http://www.nav.no/binary/1073748592/file

(17) Norwegian Insurance Administration. [Yearbook of National Insurance Statistics 2003 (In Norwegian)]. Oslo: NIA; 2003.

(18) Alexanderson K, Hensing G. More and better research needed on sickness absence. Scand J Public Health 2004;32(5):321-3.

(19) Kendrick T, Burns T, Freeling P. Randomised controlled trial of teaching general practitioners to carry out structured assessments of their long term mentally ill patients. BMJ 1995 Jul 8;311(6997):93-8.

(20) Scheel IB. The effects of promoting active sick leave for back pain patients. PhD [dissertation]. Oslo: University of Oslo; 2004.

(21) van der Klink JJ, Blonk RW, Schene AH, van Dijk FJ. Reducing long term sickness absence by an activating intervention in adjustment disorders: a cluster randomised controlled design. Occup Environ Med 2003 Jun;60(6):429-37.

(22) Bakker IM, Terluin B, van Marwijk HW, van der Windt DA, Rijmen F, van MW, et al. A cluster-randomised trial evaluating an intervention for patients with stress-related mental disorders and sick leave in primary care. PLoS Clin Trials 2007;2(6):e26.

(23) Becker A, Leonhardt C, Kochen MM, Keller S, Wegscheider K, Baum E, et al. Effects of two guideline implementation strategies on patient outcomes in primary care: a cluster randomized controlled trial. Spine 2008 Mar 1;33(5):473-80.

(24) Drøpping JA, Midtsundstad T. [Research and reporting relevant for The tripartite agreement on a more inclusive workplace (In Norwegian)]. Fafo-report 417 ed. 2003.

(25) Eklund M, von Granotz H, Marklund S. [Part-time sick leave - individual, workplace and health (In Swedish)]. In: Hogstedt C, Bjurvald M, Marklund S, Palmer E, Theorell T, editors. [The high sickness absence - truth and consequence (In Swedish)].Stockholm: Statens Folkhälsoinstitut; 2004.

(26) Norwegian Labour and Welfare Administration. [GPs' sickness certification practices II: New rules and The GP Education Programme. RTV-Report 06/2006 (In Norwegian)]. 2006.

(27) Norwegian Labour and Welfare Administration. [Persons receiving social security benefits and other arrangements administrated by the Norwegian Labour and Welfare Administration. 1997-2006 (In Norwegian)]. [Online] 2007 November 6 [cited 2008 May 2];Available from: URL: http://www.nav.no/805365427.cms

(28) Waddell G, Aylward M, Sawney P. Back pain, incapacity for work and social security benefits. An international literature review and analysis. London: Royal Society of Medicine Press; 2002.

(29) Sieurin L, Josephson M, Vingard E. Positive and negative consequences of sick leave for the individual, with special focus on part-time sick leave. Scand J Public Health 2009 Jan;37(1):50-6.

Table I. Sample characteristics of randomised GPs and their patients with long-term sick leave, and corresponding national data in 2005.

	Intervention group		Control group		General practice, Norway[1]	
Characteristics of the GPs	**n=23**	**%**	**n=29**	**%**	**n=3757**	**%**
Females	8	34.8	11	38.0	1145	30.5
Males	15	65.2	18	62.1	2612	69.5
Specialist in Family Medicine	16	69.6	24	82.8	2217	59.0
	mean	**sd**	**mean**	**sd**	**mean**	**sd**
Age, y	49.3	10.4	49.5	8.7	47.9	
Weekly working hours, h	37.5	4.7	41.3	8.5	-	
Daily consultations, n	21.8	7.2	21.0	4.8	-	
List size, n	1254.1	397.4	1309.8	210.0	1189.0	
Characteristics of the patients with long-term sick leave during the intervention period[2]	**n=939**	**%**	**n=1231**	**%**	**National data[3]**	**%**
Females	576	61.3	776	63.0	-	62.5
Males	363	38.7	455	37.0	-	37.5
Severe disease	20	2.1	33	2.7	-	-
	mean	**sd**	**mean**	**sd**	**mean**	**sd**
Mean age, y	43.7	11.8	44.2	11.5	42.0	-

[1] Numbers from The Norwegian Labour and Welfare Administration (16)

[2] Date extracted from the register of The Norwegian Labour and Welfare Administration

[3] Numbers from The Norwegian Insurance Administration (17)

Table II. Associations between intervention and duration of patient sick leave episodes analysed by Cox proportional hazards survival analysis adjusted for 52 GP clusters, age, gender, and diagnosis (intervention period, n=2170)

Variables		Hazard ratio	p-value	95% confidence interval
Group	Control group	1.0		
	Intervention group	0.89	0.071	(0.79, 1.01)
GP gender	Male	1.0		
	Female	0.97	0.651	(0.84, 1.11)
GP age		1.00	0.142	(0.99, 1.00)
Patient gender	Male	1.0		
	Female	1.13	0.038	(1.01, 1.27)
Patient age		0.99	<0.001	(0.98, 0.99
Diagnosis	Less severe disease	1.0		
	Severe disease	0.53	0.001	(0.36, 0.77)
Group*GP gender		1.27	0.033	(1.02, 1.58)

Table III. Associations between intervention and part-time sick leave analysed by binary response two-level regression model adjusted for age, gender, and diagnosis (intervention period, n=2170)

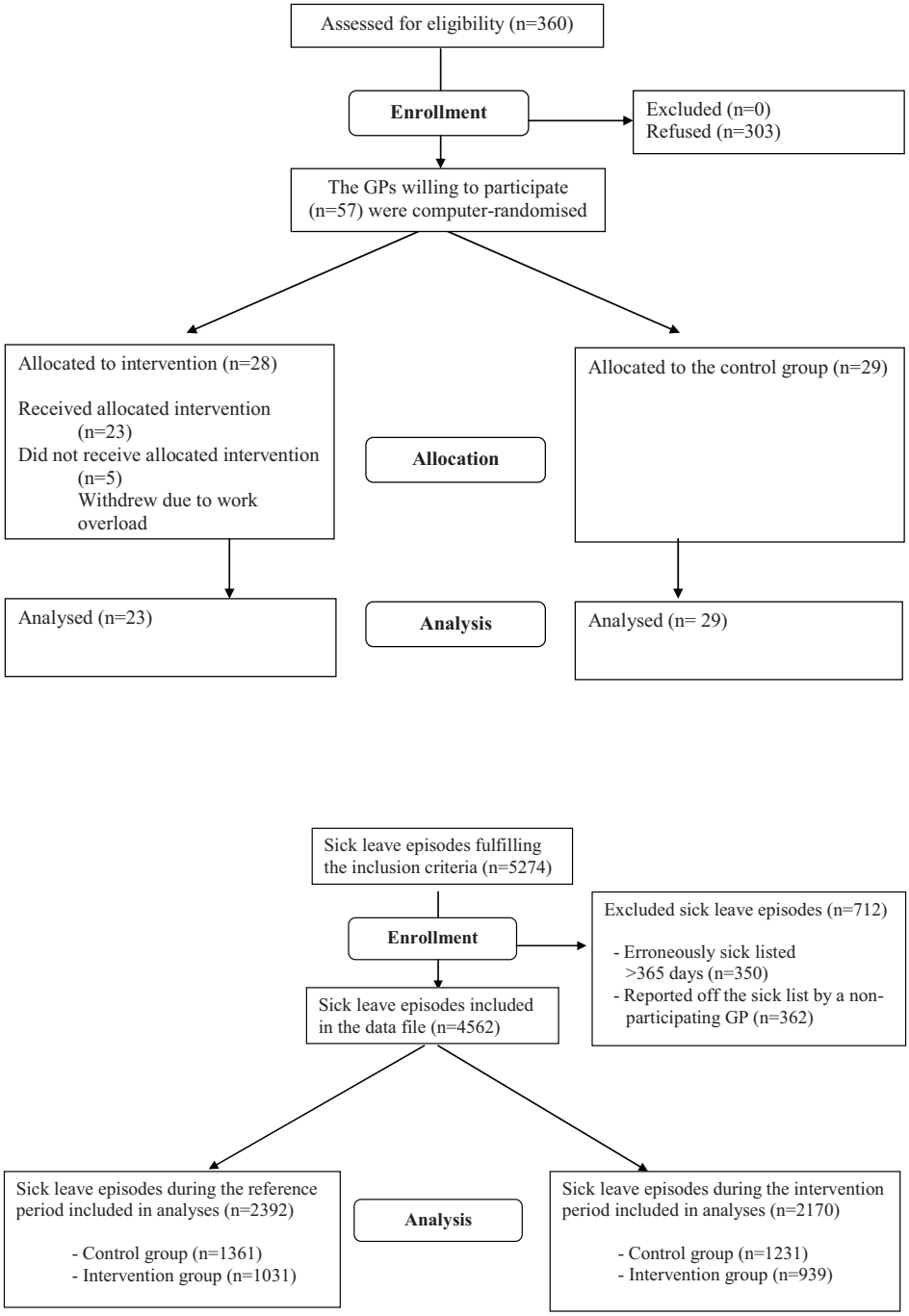
Variables		Odds ratio	p-value	95% confidence interval
Group	Control group	1.0		
	Intervention group	1.33	0.015	(1.06, 1.68)
GP gender	Male	1.0		
	Female	1.38	0.004	(1.08, 1.75)
GP age		0.99	0.120	(0.98, 1,00)
Patient gender	Male	1.0		
	Female	1.41	0.001	(1.15, 1.73)
Patient age		1.02	<0.001	(1.01, 1.02)
Diagnosis	Less severe disease	1.0		
	Severe disease	0.37	0.001	(0.21, 0.65)

Table IV. Associations between intervention and active sick leave analysed by Cox proportional hazards survival analysis adjusted for 52 GP clusters, age, gender, and diagnosis (intervention period, n=2170)

Variables		Hazard ratio	p-value	95% confidence interval
Group	Control group	1.0		
	Intervention group	0.65	0.041	(0.43, 0.98)
GP gender	Male	1.0		
	Female	1.07	0.695	(0.76, 1.52)
GP age		1.00	1.000	(0.97, 1.02)
Patient gender	Male	1.0		
	Female	0.77	0.138	(0.54, 1.09)
Patient age		0.99	0.202	(0.98, 1.01)
Diagnosis	Less serious disease	1.0		
	Serious disease	1.05	0.944	(0.26, 4.27)

Figure caption

FIGURE 1. Flow charts of GP participants and patient sick leave episodes through trial



Appendix

1: Norwegian Function Assessment Scale (five-point version) and one work ability item

2: Work Description Form

3: Key questions

4: The Function Assessment Report

5: Background questionnaire

6: Main questionnaire

7: Evaluation score-sheet (general practitioners)

8: Evaluation score-sheet (patients)

Norsk Funksjonsskjema

Har du av helsemessige grunner hatt vansker med å utføre følgende aktiviteter i løpet av **den siste uken?** Sett et kryss i den ruten som du synes passer best ved hvert spørsmål. Passer ikke spørsmålene riktig i din situasjon, ønsker vi at du likevel svarer så godt du kan. Hvis et spørsmål ikke er aktuelt for deg, f.eks fordi du aldri kjører bil, kan du sette en strek over spørsmålet.

Har du hatt vansker med å utføre følgende aktiviteter den siste uken:

Gå/stå	**Ingen vansker**	**Lite vansker**	**Middels vansker**	**Mye vansker**	**Kan ikke**
Stå	☐	☐	☐	☐	☐
Gå mindre enn 1 kilometer på flat mark	☐	☐	☐	☐	☐
Gå mer enn 1 kilometer på flat mark	☐	☐	☐	☐	☐
Gå på skiftende underlag	☐	☐	☐	☐	☐
Gå i trapper	☐	☐	☐	☐	☐
Handle dagligvarer	☐	☐	☐	☐	☐
Ta på sko og strømper	☐	☐	☐	☐	☐

Holde/plukke	**Ingen vansker**	**Lite vansker**	**Middels vansker**	**Mye vansker**	**Kan ikke**
Plukke opp en mynt fra et bord med fingrene	☐	☐	☐	☐	☐
Holde og styre et ratt med hendene	☐	☐	☐	☐	☐
Kjøre bil	☐	☐	☐	☐	☐
Lage mat	☐	☐	☐	☐	☐
Skrive	☐	☐	☐	☐	☐
Utføre vanlige oppgaver alene	☐	☐	☐	☐	☐
Drive med dine fritidsaktiviteter	☐	☐	☐	☐	☐
Kle av og på deg	☐	☐	☐	☐	☐

Løfte/bære	**Ingen vansker**	**Lite vansker**	**Middels vansker**	**Mye vansker**	**Kan ikke**
Løfte en tom bruskasse fra gulvet	☐	☐	☐	☐	☐
Bære handleposer i hendene	☐	☐	☐	☐	☐
Bære en liten sekk på skuldrene eller ryggen	☐	☐	☐	☐	☐
Skyve og dra med armene	☐	☐	☐	☐	☐
Gjøre vanlig rengjøring	☐	☐	☐	☐	☐
Gjøre klesvask	☐	☐	☐	☐	☐

Har du hatt vansker med å utføre følgende aktiviteter den siste uken:

Sitte	Ingen vansker	Lite vansker	Middels vansker	Mye vansker	Kan ikke
Sitte på en kjøkkenstol	☐	☐	☐	☐	☐
Bruke bil som passasjer	☐	☐	☐	☐	☐
Bruke kollektivtransport som passasjer	☐	☐	☐	☐	☐

Mestre	Ingen vansker	Lite vansker	Middels vansker	Mye vansker	Kan ikke
Være oppmerksom og konsentrert	☐	☐	☐	☐	☐
Arbeide i gruppe	☐	☐	☐	☐	☐
Rettlede andre i deres aktiviteter	☐	☐	☐	☐	☐
Mestre ansvar i dagliglivet	☐	☐	☐	☐	☐
Mestre dagliglivets påkjenninger og belastninger	☐	☐	☐	☐	☐
Takle kritikk	☐	☐	☐	☐	☐
Styre sinne og aggresjon	☐	☐	☐	☐	☐

Samhandling/kommunikasjon	Ingen vansker	Lite vansker	Middels vansker	Mye vansker	Kan ikke
Huske	☐	☐	☐	☐	☐
Oppfatte muntlige beskjeder	☐	☐	☐	☐	☐
Oppfatte skriftlige beskjeder	☐	☐	☐	☐	☐
Snakke	☐	☐	☐	☐	☐
Delta i samtale med flere personer	☐	☐	☐	☐	☐
Bruke telefon	☐	☐	☐	☐	☐

Sanser	Ingen vansker	Lite vansker	Middels vansker	Mye vansker	Kan ikke
Se på fjernsyn	☐	☐	☐	☐	☐
Lytte til radio	☐	☐	☐	☐	☐

Arbeidsevne

I hvilken grad er din evne til å utføre ditt vanlige arbeid nedsatt akkurat i dag?

☐ **Ubetydelig nedsatt** ☐ **Ikke særlig nedsatt** ☐ **Middels nedsatt** ☐ **Mye nedsatt** ☐ **Svært mye nedsatt**

Arbeidsbeskrivelse

Type arbeid: ……

Hvor lenge har du vært ansatt? ☐ Under 1 år ☐ 1-5 år ☐ Over 5 år

Yrkesstatus: ☐ I arbeid ☐ Arbeidsledig ☐ Attføring/rehab.

Arbeider du heltid eller deltid? ☐ Heltid ☐ Deltid

Nevn 3 positive sider ved arbeidet:

1. _____ 2._____ 3._____

Opplever du arbeidet som fysisk belastende? ☐ Nei ☐ Ja

Hvis JA, sett ett eller flere kryss:

- ☐ Mye stillesitting
- ☐ Står stille
- ☐ Går mye
- ☐ Står på kne eller sitter på huk
- ☐ Arbeider med armene løftet/fremstrakt
- ☐ Løfter mye tungt
- ☐ Utfører nøyaktige bevegelser med hendene
- ☐ Utfører samme bevegelser mange ganger i minuttet
- ☐ Arbeider på underlag/med verktøy som vibrerer
- ☐ Må holde samme arbeidsstilling lenge
- ☐ Har tungt arbeid
- ☐ Annet: ……

Opplever du arbeidet som mentalt belastende? ☐ Nei ☐ Ja

Hvis JA, sett ett eller flere kryss:

- ☐ Må være oppmerksom og konsentrert
- ☐ Må takle følelser
- ☐ Må ha god hukommelse
- ☐ Annet: ……..
- ☐ Må ha evne til nytenking
- ☐ Arbeider sammen med kollegaer om oppgaver
- ☐ Direkte kontakt med klienter, kunder eller elever

Opplever du arbeidsorganiseringen som belastende? ☐ Nei ☐ Ja

Hvis JA, sett ett eller flere kryss:

- ☐ Har skiftarbeid
- ☐ Arbeider på akkord/provisjon
- ☐ Har sesongintensivt arbeid
- ☐ Har lederansvar
- ☐ Har for mye å gjøre
- ☐ Har for stort ansvar
- ☐ Annet: ……..
- ☐ Uklart hva som forventes på jobben
- ☐ Kan ikke bestemme arbeidstempo selv
- ☐ Kan ikke bestemme selv når jeg kan ta pauser
- ☐ Får ikke hjelp med de tyngste oppgavene
- ☐ Får lite støtte og hjelp av overordnede
- ☐ Synes ikke min arbeidsinnsats blir verdsatt

Nøkkelspørsmål

Spørsmålene stilles muntlig til pasienten etter gjennomgang av Norsk Funksjonsskjema og Arbeidsbeskrivelse (se veiledning). Gjør pasienten oppmerksom på at svarene på disse spørsmålene ikke formidles videre.

Hvor mye lenger tror du at du vil være sykmeldt?

Klarer du å jobbe litt (evt. bare noen timer per uke) i den jobben du har nå?
Svarer pasienten ja på dette spørsmålet, bør graderte sykepenger (delvis sykmelding) diskuteres. Det vil være gunstig for å holde kontakten med arbeidsplassen.

Tenker du for øyeblikket på å få annet arbeid?
Svarer pasienten ja på dette spørsmålet, bør det diskuteres hvordan han eller hun tenker seg overgangen til annen jobb. Kan det være aktuelt med yrkesrettet attføring?

Alle forhold tatt i betrakting, hvor viktig er jobben for deg?
Om jobben er viktig for pasienten, bør rehabiliteringsmulighetene diskuteres inngående.

Har vanskelig arbeidsmiljø eller konflikter på arbeidet medvirket til fraværet?
Drøft med pasienten hvordan dette skal tas opp. Kontakte bedriftshelsetjenesten?

Hvordan opplever du kravene som stilles til deg utenom jobben, dvs. hjemme, i familien og på fritiden?
Dersom det er andre krav utenfor jobben som er belastende, bør legen prøve å klarlegge dette.

Funksjonsrapport

Medisinsk funksjonsvurdering etter forespørsel fra trygdeetat eller arbeidsgiver.

Skjemaet fylles ut av behandlende lege i samtale med den sykmeldte på bakgrunn av Norsk Funksjonsskjema og Arbeidsbeskrivelse.

Sykmeldtes navn: ……. Fødselsdato ☐☐☐☐☐☐

Stilling: ……... Antall uker sykmeldt: ☐☐ Uker

Legens navn:…… Samtaledato ☐☐☐☐☐☐

1. Funksjonsevne (jf. Norsk Funksjonsskjema)

Kryss av der den sykmeldte har angitt ressurser eller svikt.

FUNKSJONSOMRÅDE	Ressurser	Svikt
Gå/stå		
Holde/plukke		
Løfte/bære		
Sitte		
Mestre		
Samhandling		
Sanser		

Kommentarer:

Arbeidsevne (jf. Norsk Funksjonsskjema)

Sett ett kryss for hvilken grad av nedsatt arbeidsevne den sykmeldte har angitt.

Ubetydelig nedsatt	Ikke særlig nedsatt	Middels nedsatt	Mye nedsatt	Svært mye nedsatt

2. Arbeidsbeskrivelse (jf. Arbeidsbeskrivelse)

Har den sykmeldte angitt at arbeidet er belastende? Skriv i kolonnen til høyre hva som oppleves belastende.

Arbeidsbeskrivelse	Nei	Ja	Hvis ja, hva oppleves belastende?
Fysisk belastende			
Mentalt belastende			
Belastende elementer ved arbeidsorganiseringen			

3. Legens vurdering av ressurser og funksjonssvikt i forhold til arbeidsoppgaver

Hvilke ressurser har den sykmeldte som kan utnyttes i tilbakeføringen til arbeid?

Har den sykmeldte spesielle behov, f.eks for pauser og hvile?

Kan den sykmeldte arbeide deltid? ☐ Ja ☐ Nei

4. Medisinsk behandling

Vil pågående eller planlagt behandling påvirke den sykmeldtes funksjonsevne? Er behandlingen til hinder for aktivitet? Når avsluttes behandlingen?

5. Skånebehov

Angi bestemte situasjoner eller ytre forhold som den sykmeldte av medisinske grunner skal unngå, som f.eks løfte/bære, jobbe med armene hevet eller kunde-/elevkontakt/klientbehandling.

6. Forslag til tiltak på arbeidsplassen

Vil tilrettelegging eller hjelpemidler gjøre det lettere for den sykmeldte å komme tilbake på jobb?

☐ Antagelig ja ☐ Vet ikke ☐ Antagelig nei

Konkrete forslag til tilrettelegging på arbeidsplassen. Hva kan gjøre tilbakegangen til arbeidet lettere?

7. Kommentarer

.....	.....	……..
Dato	Legens underskrift	Den sykmeldtes underskrift

UNIVERSITETET I OSLO

"Funksjonsvurderinger ved langvarig sykefravær"

LEGESKJEMA

1. Kjønn:	☐ Mann ☐ Kvinne
2. Fødselsår:	19 ____
3. Eksamensår:	19 ____
4. Antall år i nåværende praksis:	____ år
5. Er du spesialist i allmennmedisin og/eller samfunnsmedisin?	☐ Nei ☐ Ja, i allmennmedisin ☐ Ja, i samfunnsmedisin ☐ Ja, i både allmenn- og samfunnsmedisin
6. Ordinært antall arbeidstimer pr uke:	___ timer pr.uke
7. Type praksis:	☐ Solopraksis ☐ Gruppepraksis
8. Praksislokalisasjon:	☐ By med mer enn 15 000 innbyggere ☐ By med mindre enn 15 000 innbyggere ☐ Tettsted ☐ Grisgrendt strøk
9. Oppgjørsordning:	☐ Fastlønn ☐ Driftstilskudd ☐ Ingen offentlig finansiering ☐ Annet, beskriv:_____
10. Gjennomsnittlig antall pasientkonsultasjoner pr. dag:	_____ pr. dag
11. Antall pasienter på "fastlegelista":	____ pasienter
12. Har du deltatt på Fagutviklings-programmet "Legerollen i det inkluderende arbeidsliv" i regi av Trygdeetaten og Aplf?	☐ Nei ☐ Ja

UNIVERSITETET I OSLO

"Funksjonsvurderinger ved langvarig sykefravær"

SPØRRESKJEMA NR. 1

Med arbeidsrelatert aktivitet mener vi graderte sykepenger, aktiv sykmelding, yrkesrettet attføring, tilretteleggingstilskudd eller reisetilskudd. Langtidssykefravær regnes her som sykefravær over 8 uker.

Sett ett kryss for hver linje.

Hvordan bedømmer du dine kunnskaper om:	**Svært gode**	**Gode**	**Middels**	**Dårlige**	**Svært dårlige**
1. Funksjonsvurderinger av langtidssykmeldte pasienter	☐	☐	☐	☐	☐
2. Arbeids-/yrkesmessig rehabilitering av langtidssyk-meldte pasienter	☐	☐	☐	☐	☐
3. Arbeidsrelatert aktivitet for langtidssykmeldte pasienter	☐	☐	☐	☐	☐

Ta stilling til følgende utsagn:	**Helt enig**	**Delvis enig**	**Verken enig eller uenig**	**Delvis uenig**	**Helt uenig**
4. Funksjonsvurderinger er viktig for å få den langtidssykmeldte pasienten tilbake i arbeid	☐	☐	☐	☐	☐
5. Arbeids-/yrkesmessig rehabilitering er viktig for å få den langtidssykmeldte pasienten tilbake i arbeid	☐	☐	☐	☐	☐
6. Arbeidsrelatert aktivitet er viktig for å få den langtids-sykmeldte pasienten tilbake i arbeid	☐	☐	☐	☐	☐
7. Funksjonsvurderinger er viktig for trygdeetaten i forhold til å kunne bistå langtidssykmeldte med å komme tilbake i arbeid	☐	☐	☐	☐	☐
8. Funksjonsvurderinger er viktig for arbeidsgiver i forhold til å kunne tilrettelegge arbeidet for den langtidssykmeldte	☐	☐	☐	☐	☐
9. Jeg opplever det som meningsfylt å gjøre funksjonsvurderinger på langtidssykmeldte pasienter	☐	☐	☐	☐	☐

SNU ARKET

Ta stilling til følgende utsagn:	Helt enig	Delvis enig	Verken enig eller uenig	Delvis uenig	Helt uenig
10. Jeg opplever det som bortkastet tid å gjøre funksjonsvurderinger på langtidssykmeldte pasienter	☐	☐	☐	☐	☐
11. Jeg anser meg selv som godt rustet til å gjøre funksjonsvurderinger av mine langtidssykmeldte pasienter	☐	☐	☐	☐	☐
12. Funksjonsvurderinger av langtidssykmeldte pasienter bør foretas av andre yrkesgrupper enn leger	☐	☐	☐	☐	☐
13. Jeg er overbevist om at jeg kan foreta en funksjonsvurdering av en langtidssykmeldt pasient dersom pasientens arbeidsgiver eller trygdekontor ber om det	☐	☐	☐	☐	☐
14. Jeg er overbevist om at jeg vet hvordan jeg kan finne fram til arbeidsrelevante ressurser hos en langtidssykmeldt pasient	☐	☐	☐	☐	☐

Hvordan bedømmer du dine kunnskaper om:	Svært gode	Gode	Middels	Dårlige	Svært dårlige
15. Arbeidsplassene til dine langtidssykmeldte pasienter	☐	☐	☐	☐	☐
16. Arbeidsoppgavene til dine langtidssykmeldte pasienter	☐	☐	☐	☐	☐
17. Fysiske belastninger på arbeidsplassene til dine langtidssykmeldte pasienter	☐	☐	☐	☐	☐
18. Mentale belastninger på arbeidsplassene til dine langtidssykmeldte pasienter	☐	☐	☐	☐	☐
19. Belastninger knyttet til organiseringen av arbeidet på arbeidsplassene til dine langtidssykmeldte pasienter	☐	☐	☐	☐	☐

UNIVERSITETET I OSLO

SPØRRESKJEMA **TIL LEGEN** ETTER KONSULTASJONEN

Dato for konsultasjon: ☐☐☐☐☐☐

Pasientens kjønn: Mann ☐ Kvinne ☐

Pasientens fødselsår: ☐☐☐☐

Hvilket kjennskap	**Intet**	**Noe**	**Godt**	**Meget godt**	**Svært godt**
1. hadde du til denne pasientens arbeidsplass, arbeids-oppgaver og arbeidskrav før konsultasjonen i dag?	☐	☐	☐	☐	☐
2. hadde du til denne pasientens fysiske belastninger i arbeidet før konsultasjonen i dag?	☐	☐	☐	☐	☐
3. hadde du til denne pasientens mentale belastninger i arbeidet før konsultasjonen i dag?	☐	☐	☐	☐	☐
4. hadde du til denne pasientens belastninger på grunn av arbeidsorganiseringen før konsultasjonen i dag?	☐	☐	☐	☐	☐
5. har du til denne pasientens arbeidsplass, arbeids-oppgaver og arbeidskrav etter konsultasjonen i dag?	☐	☐	☐	☐	☐
6. har du til denne pasientens fysiske belastninger i arbeidet etter konsultasjonen i dag?	☐	☐	☐	☐	☐
7. har du til denne pasientens mentale belastninger i arbeidet etter konsultasjonen i dag?	☐	☐	☐	☐	☐
8. har du til denne pasientens belastninger på grunn av arbeidsorganiseringen etter konsultasjonen i dag?	☐	☐	☐	☐	☐

UNIVERSITETET I OSLO

SPØRRESKJEMA **TIL PASIENTEN** ETTER KONSULTASJONEN

Dato for konsultasjon: ☐☐☐☐☐☐

Kjønn: Mann ☐ Kvinne ☐

Fødselsår: ☐☐☐☐

Hvilket kjennskap tror du din lege	**Intet**	**Noe**	**Godt**	**Meget godt**	**Svært godt**
1. hadde til din arbeidsplass, arbeidsoppgaver og arbeidskrav før konsultasjonen i dag?	☐	☐	☐	☐	☐
2. hadde til fysiske belastninger i arbeidet ditt før konsultasjonen i dag?	☐	☐	☐	☐	☐
3. hadde til mentale belastninger i arbeidet ditt før konsultasjonen i dag?	☐	☐	☐	☐	☐
4. hadde til belastninger på grunn av arbeidsorganiseringen i arbeidet ditt før konsultasjonen i dag?	☐	☐	☐	☐	☐
5. har til din arbeidsplass, arbeidsoppgaver og arbeidskrav etter konsultasjonen i dag?	☐	☐	☐	☐	☐
6. har til fysiske belastninger i arbeidet ditt etter konsultasjonen i dag?	☐	☐	☐	☐	☐
7. har til mentale belastninger i arbeidet ditt etter konsultasjonen i dag?	☐	☐	☐	☐	☐
8. har til belastninger på grunn av arbeidsorganiseringen i arbeidet ditt etter konsultasjonen i dag?	☐	☐	☐	☐	☐