UNIVERSITY OF TROMSØ UIT

FACULTY OF HEALTH SCIENCES
REGIONAL CENTER FOR CHILD AND YOUTH MENTAL HEALTH AND CHILD WELFARE

Assignment of Mental Health Diagnoses and Severity

Effectiveness and Reliability of Online Standardized Assessment Instruments

Per Håkan Brøndbo

A dissertation for the degree of
Philosophiae Doctor

Mai 2012

Assignment of Mental Health Diagnoses and Severity: Effectiveness and Reliability of Online Standardized Assessment Instruments

Per Håkan Brøndbo

A dissertation for the degree of Philosophiae Doctor (Ph.D.)

2012

Regional Centre for Child and Youth Mental Health and Child Welfare

Faculty of Health Sciences

University of Tromsø

Table of contents

Acknowledgements

The work described in the present dissertation was carried out between January 2009 and March 2012. I would like to thank the Northern Norway Regional Health Authority, the University Hospital of North-Norway and the University of Tromsø which funded the Child and Adolescent Mental Health Services North Study. I would also like to thank the Regional Center for Child and Youth Mental Health and Child Welfare, University of Tromsø, for providing financial support for the training of raters. I would also like to thank all the participating patients, parents and teachers, in addition to all mental health workers at the Child and Adolescent Mental Health Outpatient Clinics in Tromsø and Alta. There would be no study or results to present without you. Thank you all!

Furthermore I would like to thank my Ph.D. supervisor, Siv Kvernmo, who is the former head of the Department of Child and Adolescent Mental Health and also the Head of the CAMHS North Study. You have always encouraged my work, and have generously shared your knowledge and network with me. Your enthusiasm and continuous stream of new ideas and project plans are an inspiration, and give me something to strive for. Thank you, and I look forward to our next project together!

My supervisor, Monica Martinussen, is definitively 'one of a kind'. Thank you for always being there, no matter where your physical location. Your methodological insight, quick responses, and confident supervision in combination with all your enthusiastic comments on my work have been invaluable to me. Thank you, and I really hope that we will continue to collaborate!

My co-authors, Einar Heiervang, Mads Eriksen, Therese Fjeldmo Moe, Guri Sæther and Bjørn Helge Handegård, have helped me in different ways by reading, writing and providing constructive input to the articles published in the framework of this dissertation. For that I thank you!

Børge Mathiassen, my friend, colleague and now head of the Department of Child and Adolescent Mental Health, has been my co-author from start to finish. We have shared many discussions, frustrations and also good meals! Your knowledge, pragmatic views and capacity to bear enormous workloads have been inspiring, as has your conviction that all our findings were worthy of publication even at an early stage. I have really appreciated working with you, and look forward to further collaboration on new projects.

I would also like to thank Trudy K. Perdrix-Thoma of Professional Standards Editing. Your excellent language review and comments on what aspects of my texts needed clarification have not only lifted the quality of my manuscripts, but also taught me some invaluable lessons that I can integrate into my future writing.

Last and most warmly I want to thank my family and friends for all their support and encouragement during these years. Kari, you have been an inspiring, insightful and constructive discussion partner, in addition to gently pushing me towards completing this dissertation. You and Sondre are my loved ones and make my life complete in all ways!

To all whom I may have forgotten, please blame it on my head not my heart!

List of research papers

1. Brondbo PH, Mathiassen B, Martinussen M, Heiervang E, Eriksen M, Kvernmo S. Agreement on web-based diagnoses and severity of mental health problems in Norwegian child and adolescent mental health services. Clin Pract Epidemiol Ment Health 2012; 8:16-21.
2. Brondbo PH, Mathiassen B, Martinussen M, Heiervang E, Eriksen M, Moe T, Saether G, Kvernmo S. The strengths and difficulties questionnaire as a screening instrument for Norwegian child and adolescent mental health services, application of UK scoring algorithms. Child Adolesc Psychiatry and Ment Health 2011; 5:32.
3. Brondbo PH, Mathiassen B, Martinussen M, Handegard BH, Kvernmo S. Agreement on diagnoses and severity of mental health problems between a research and a naturalistic clinical setting. J Telemed Telecare submitted.

Abstract

During the last decade, child psychiatry has been a focus of the Norwegian government's plan to improve mental health care. Psychiatric assessment of children and adolescents is more complex than that of adults, and rating scales and diagnostic instruments have become increasingly important tools in both research and clinical practice.

This dissertation investigates standardized assessment instruments used in routine clinical practice to assign diagnoses and severity of mental health problems. Both the reliability and the validity of some instruments were examined, and a main focus was the clinical usefulness of these instruments and their potential for more effective use of limited clinical resources.

In paper 1 the agreement between diagnoses and severity ratings assigned by clinical specialists who were trained Child and Adolescent Mental Health Service (CAMHS) researchers examined. Information on 100 youths was obtained from multiple informants through a web-based Development and Well-Being Assessment (DAWBA). Based on this information, four experienced clinicians independently diagnosed (according to the International Classification of Diseases Revision 10) and rated the severity of mental health problems according to the Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA) and the Children's Global Assessment Scale (C-GAS). Agreement for diagnosis was κ =0.69-0.82. Intra-class correlation for single measures was 0.78 for HoNOSCA and 0.74 for C-GAS, and 0.93 and 0.92, respectively for average measures. Information obtained with the online DAWBA may be a sound basis on which to establish reliable clinical diagnoses and severity ratings for common mental health disorders in a clinical setting. A clinical practice that includes systematic, multiple independent assignments of diagnosis and severity, is preferable due to the resulting improved reliability of the severity ratings.

In paper 2 the application of specific scoring algorithms for the Strengths and Difficulties Questionnaire (SDQ) was examined. Could available online norms be useful in screening for mental health disorders among children and adolescents in the CAMHS North Study? A total of 286 outpatients, aged 5 to 18 years, were assigned diagnoses based on the DAWBA. The main diagnostic groups (emotional, hyperactivity, conduct and other disorders) were then compared to the SDQ scoring algorithms using two dichotomisation levels: 'possible' and 'probable' levels. Sensitivity for the diagnostic categories included was 0.47-0.85 ('probable' dichotomisation level) and 0.81-1.00 ('possible' dichotomisation level). Specificity was 0.52-0.87 ('probable' level) and 0.24-0.58 ('possible' level). The discriminative ability, as measured by OR^D, was in the interval for potentially useful tests for hyperactivity disorders and conduct disorders when dichotomised on the 'possible' level, but outside the interval for potentially useful tests for all diagnostic categories when dichotomised on the most common used 'probable' level. In conclusion, the ability of the SDQ to detect mental health disorders among patients referred to CAMHS is not sufficient for clinical purposes. When used as a screening instrument to determine whether further evaluation is warranted in a clinical CAMHS sample, the SDQ seems best suited to identify children and adolescents who do not require further psychiatric evaluation, although this also is problematic from a clinical point of view.

In paper 3 the agreement between diagnoses and severity assigned by clinical specialists trained as CAMHS researchers, based only on DAWBA information collected online, and the routine clinical assignments by CAMHS clinicians was examined. Routine clinical assignment of diagnoses was compared to online clinical assignment of diagnoses for 286 patients from the CAMHS North study. Chi square analysis, kappa statistics and multinomial logistic regression analyses were performed. Raw agreement for diagnostic categories varied between 74% and 90%, resulting in kappa values of 0.41-0.49. The final multinomial regression models were

significant. Agreement on mental health diagnoses can be fair when online clinical assignments and routine clinical assignments of mental health diagnoses are compared. This may be sufficient to replace the routine clinical assignment of diagnoses with an online clinical assignment in order to save time and resources. We also examined factors contributing to agreement or disagreement on the diagnoses. Age, gender and number of informants significantly contributed to the explanation of agreement and disagreement for 'emotional diagnosis' and 'hyperkinetic/conduct diagnosis'. However, the changes in odds were small in magnitude and the factors probably do not consistently contribute to the understanding of agreement or disagreement in any clinically meaningful way.

Lastly, implications for further research on reliable and effective assessment methods are discussed.

Introduction

During the last 10 years, child psychiatry has been a focus of the Norwegian government's plan to improve mental health care, the aim being to raise the bar for competence and research in of the field; to increase the number of mental health workers and the availability of mental health services for children [1]. This plan has focused broadly on evidence-based methods, i.e., methods with a proven positive effect, but most attention to-date has been given to evidence-based treatment. Less focus has been placed on evidence-based assessment and what constitutes an assessment instrument that is "good enough" for use in both research and clinical settings [2]. Psychiatric assessment of children and adolescents is more complex than that of adults, due to the necessity to involve both the family and the school. Although many studies have shown that clinical assessment is not better than actuarial algorithms [3, 4], clinicians still tend to use unstructured clinical interviews when assessing children and adolescents for mental health problems [5]. However, a recent study showed better diagnostic agreement and accuracy among clinicians after they attended a brief training session in the use of a structured instrument designed to effectively combine multiple sources of information. In addition, a majority of these clinicians reported a positive attitude towards the use of this instrument in routine clinical practice [6, 7]. Both society and the individuals seeking mental health services benefit if these services possess effective assessment and treatment methods [8-10]. Various measures (i.e., average time on waiting list, number of days to send out a report, number of patients with a recorded diagnosis) have primarily been used to evaluate the outpatient clinics themselves, but diagnostic practices and efficacy of various measures used in the diagnostic process have not been sufficiently studied [11].

Child and Adolescent Mental Health Services in Northern Norway

The Child and Adolescent Mental Health Services North (CAMHS North) Study was carried out in the northern part of Norway; the main goals were to evaluate clinical procedures and treatment, to investigate factors that may affect the waiting list, to evaluate examination and treatment time, to implement and validate structured instruments, and to investigate user satisfaction. Northern Norway, located at the very northern periphery of Europe, covers about 35% of the Norwegian mainland, but is inhabited by only about 10% of the Norwegian population. A recent study showed good coping skills regarding help seeking for both physical and mental health among adolescents in Northern Norway [12]. However, in this region CAMHS coverage, the stipulated needs, which are based on socio-economic variables, and the actual demand for CAMHS are far higher than the national average [13]. This may be related to geographic variations, as well as the organization and scope of municipal services dedicated to children and adolescents, but may also be related to real differences in mental illness across regions, although youths in Northern Norway reported lower or equal rates of behavioral/emotional problems compared to a nationally representative sample [14].

Most child and adolescent mental health patients are treated in outpatient clinics. Regional health authorities are responsible for patient-oriented research, research training, dissemination of research results and implementation of useful research conclusions, as well as to provide services of high professional quality through continuous quality improvement. Health services should always act in accordance with good professional practice and current regulations, including the definition of evidence-based practice in psychology as “the integration of the best available research with clinical expertise in the context of patient characteristics, culture, and preferences” (273) [15]. All patients of outpatient clinics should be examined for potential problems, manifestations of symptoms , functioning in daily life, care situation, educational situation, risk

factors, resources and patient's/parent's wishes and expectations. Systematic use of structured diagnostic interviews, questionnaires and standardized assessment instruments is recommended, so that professionals can methodically ask about and consider the full spectrum of a patient's symptoms and features [13]. Diagnostic assignment should be the result of an overall assessment of the patient's condition and circumstances, and should include both problems and resources. Research from New Zealand has shown that families of adult psychiatric service users have positive views of the diagnostic practice there, but suggested that more contextual issues, such as financial and family dynamic stressors, accommodation and life skills, should be taken into account. In that study, the most important element in a positive experience was how the diagnoses were communicated to patients and family members, and how these diagnoses were utilized in treatment planning [16]. However, negative experiences with diagnostic practice have been reported for families of child psychiatric service users [17, 18]. Parents' dissatisfaction was strongly associated with long delays in confirming diagnoses and a high number of professionals consulted before obtaining a diagnosis [17]. A Norwegian study found that the mean diagnosis time, from identification of hyperkinetic disorders by parents to a clinical diagnosis assigned by mental health service providers, was about 4 years [19].

Lack of 'gold standard' for mental health diagnoses

Despite advances in the classification systems, including the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV) and the International Classification of Mental and Behavioural Disorders, 10th revision (ICD-10), mental health diagnostics remain based on subjective markers such as developmental history, behavioral observations and reported difficulties in everyday life. The accuracy of diagnostic assignment depends on the clinician's capability to operationalize the criteria in the DSM-IV and ICD-10, while integrating information

from different sources and perspectives [20, 21]. No commonly accepted ‘gold standard’ is available. However, in research settings, structured interviews such as the DAWBA are often used as a ‘gold standard’ [26], while in clinical settings unstructured clinical interviews are most often used to generate diagnoses [5, 27].

The acceptance of clinician consensus diagnoses as the ‘gold standard’, is problematic, as there is no single objective feature that distinguishes any mental health diagnosis. Costello et al [20] stated that structured interviews are the closest we can come to a ‘gold standard’ for psychiatric diagnoses. Thus, a diagnostic assignment of a single clinical expert that is aided by a structured interview such as the Development and Well-Being Assessment (DAWBA) may be the best available ‘gold standard’ reference. However, the use of a single expert rating may not always be sufficient to achieve reliable diagnoses [22]. A consensus discussion provides intelligent input from several experts in order to refine the final diagnosis, and although consensus procedures are also imperfect, they will remain valuable as long as mental health diagnoses are based on the same subjective markers.

Structured diagnostic interviews versus unstructured clinical interviews

The systematic use of structured diagnostic interviews, questionnaires and standardized assessment instruments is part of routine procedure in most mental health research [23]. However many clinicians argue that this descriptive approach, which is based on the diagnostic criteria outlined in both the DSM and ICD, does not fit the clinical reality [24]. An important goal when using standardized assessment instruments in a clinical setting is to enhance the agreement and accuracy of diagnoses among clinicians with different backgrounds and levels of experience. The use of structured interviews increases the likelihood of accomplishing good inter-rater reliability,

but does not ensure it. Both the reliability and validity of mental health diagnoses in routine clinical practice has been questioned [25].

In a clinical setting, observation of the child is expected before making a diagnosis, whereas most research does not include this kind of information in the diagnostic process. This is in part due to the fact that clinical assessment is often focused on case conceptualization and tailoring treatment rather than categorization of diagnoses [28]. Agreement between diagnoses based on structured interviews and clinical diagnostic assignments was found to be low to moderate in a recent meta-analysis (overall agreement kappa [κ] = 0.15) [26]. Reducing the gap between research-derived knowledge and clinical practice in CAMHS is a challenge, but it is important to improve the rationality, efficiency and quality of service [29].

Inclusion of the diagnostic criterion *disability*, defined as impairment in one or more important areas of functioning (social, academic, occupational, etc.), has resulted in lower prevalence estimates and greater agreement on diagnoses [30, 31]. But the agreement between clinicians for common mental disorders still ranges from low to moderate [32, 33]. Clinical experience, immediate feedback on the prediction, available objective instruments to aid diagnostic accuracy, and available base rate information may improve diagnostic assignment [3, 34].

Use of structured interviews instead of unstructured clinical interviews has been shown to significantly improve diagnostic accuracy [35, 36]. Miller et al [37] found a 45.5% raw inter-rater agreement between experienced clinicians for traditional diagnostic assignment, with a κ of 0.24. The equivalent numbers for a computer-assisted structured diagnostic interview were 79.5% and 0.75, respectively, in the same study. In a review of clinical diagnoses of depression, Williams et al [38] found an inter-rater agreement for mental health care professionals ranging from $\kappa = 0.64$ to 0.93 when the diagnostic process was supplemented by semi-structured interviews. For

diagnoses assigned without the aid of this instrument, agreement was $\kappa = 0.55$ to 0.74. Foreman et al [39] found that diagnoses of mental health problems, based on information from the DAWBA, were sufficiently accurate without direct patient contact. They reported joint reliability for clinical and DAWBA-generated diagnoses ranging from $\kappa = 0.57$ to 0.76, and positive and negative predictive values (PPV, NPV) greater than 0.75 for DAWBA-generated diagnoses.

Categorical and dimensional diagnoses

The ICD system was developed due to a need to define the entire range of mental health diseases in a format fitted for statistical analysis [23]. The size and complexity of the system has increased from approximately 10 pages in the 6th edition to around 300 pages in the 10th edition [40]. The descriptive, atheoretical approach, which focused on phenomenology rather than etiology or pathogenesis, has led to improvement in the identification and treatment of mental disorders [41]. Although the improved diagnostic reliability in research after the introduction of the ICD system is well documented [42], diagnostic reliability and validity in routine clinical practice have both been questioned [25]. However, studies comparing daily behavior, impaired functioning and longitudinal outcome to both clinical, and research-generated diagnoses have indicated higher validity for the latter [43, 44]. A recent Swedish report concluded that the flora of structured and semi-structured assessment instruments used in CAMHS is rapidly growing, but that there is a lack of knowledge about the reliability and validity of such instruments. Indeed, less than one out of four of them is adapted to local or national conditions and fulfills the quality criteria for assessment instruments [45].

Despite the lack of clear boundaries between different psychiatric disorders and the fact that the validity of categorical mental health diagnoses can be questioned, a categorical, international nosology is useful for research purposes. The importance of categorical diagnoses is also obvious

in a clinical setting, where many situations are categorical by nature and where decisions about different treatment options and prognostic predictions are made, [23]. A correct diagnosis can ensure appropriate treatment at an early stage, leading to faster recovery and a shorter treatment period. Development of diagnosis-specific treatment protocols also strengthens the need for reliable and valid diagnostic procedures [32, 35].

Although the ICD-10 was designed to facilitate clinical practice, the diagnoses described therein, and their thresholds, are too complex to be useful in routine clinical practice [46]. Serious concerns about the clinical usefulness of categorical diagnoses have been raised based on the extensive use in clinical practice of 'non-specific' diagnoses, the high proportion of comorbidity, the low specificity between diagnoses and effective psychotherapy or pharmacotherapy, the numerous diagnostic distinctions with little or no clinical relevance, and the poor predictive power of treatment needs [41]. It has therefore been suggested that minimal time and expense should be dedicated to diagnostic assessment [47], making the clinical usefulness of the ICD-10 a major concern in its next revision. Indeed, in the revision the complexity of the diagnoses must be reduced, and the use of the ICD system must be simplified so that feasibility of usage in different multidisciplinary health care settings can be improved. Andrews et al [46] claimed that the main problems with the current classifications are the complexity and the lack of evidence for thresholds and exclusion criteria for many diagnoses. In addition, there is a growing agreement regarding the dimensional nature of mental health disorders in general, and for children and adolescents in particular [48, 49]. A dimensional approach to mental health disorders can accommodate the range of expressions of psychopathology in children of different ages and genders according to both character and magnitude as development progresses. In addition, neither the ICD-10, nor the DSM-IV include guidelines on how to handle discrepancies between multi-informant sources (i.e., children's, parent's and teacher's reports of mental health

problems). Such discrepancies are common [50] and can be analyzed and used in more sophisticated ways than just categorizing the child as sick or healthy [51, 52]. Kraemer [49] argued that the only situation where a dimensional diagnosis does not add quality to a categorical diagnosis is when there is no meaningful clinical variation among either positively- or negatively-diagnosed patients. A critique of the dimensional approach has been its reduced clinical usefulness [53]. Clinicians need cut-off points in order to make decisions about treatment and admission to services, and most such decisions are categorical by nature. Thus, a combination of categorical and dimensional classifications, which is the norm in most areas of medicine, may better utilize evidence-based approaches to child and adolescent psychiatry [53].

Another relevant question may be: Could a meta-structure with large clusters of diagnoses, identified by external validating factors (i.e., genetics, epidemiology, risk factors, therapeutics) improve the validity and utility of diagnostic assignment, and facilitate both research and patient care [54]? Correlations and factor analyses have shown robust, generalizable results for genetic data [55] across clinical and non-clinical samples [56], and across different countries and cultures [57]. Identification of clusters of diagnoses may enhance the clinical utility of diagnostic assignment by simplifying it, but may also benefit treatment by reducing the impact of risk factors that are characteristic of a specific cluster. Clusters may also simplify data reporting and public health planning, because 'severe mental disorders' (i.e., psychoses and neurocognitive disorders) call for different types of services and mental health professionals than do neurodevelopmental disorders (e.g., autism, mental retardation). Patients with neurodevelopmental disorders, in turn, face different challenges than those with 'common mental disorders' (i.e., emotional and externalizing disorders). In a review article, Andrews et al [54] concluded that risk factors and "clinical profile" were shared by such large groups of disorders (i.e., neurocognitive, neurodevelopmental, psychosis, emotional, externalizing, and disorders not

yet assigned) that use of those clusters could be advantageous for clinical practice, public health planning and research purposes.

The gap between treatment needs and capacity

Mental health services face the challenge of a growing trend in earlier "age of onset", greater severity, and increasing comorbidity among today's children and adolescents [10, 58, 59]. A conservative prevalence estimate of psychiatric disorders in the Norwegian child and adolescent population (3-18 years old) is about 8% based on epidemiological surveys [60]. One large study showed a prevalence of 7% among children aged 8 to 10 years [61]. It is even more common for children and adolescents to suffer psychosocial impairment due to mental health problems, with an estimated 15 to 20% of this age group being affected [60]. CAMHS in Norway are supposed to cover 5% of the child and adolescent population according to the Norwegian Health Authorities [62]. Service needs are not predicted solely by the number of children and adolescents diagnosed, but also by those who display psychosocial impairment without assigned diagnoses [63]. The gap between the prevalence/impairment estimates and CAMHS coverage highlights a very real capacity problem in the Norwegian mental health care system, which results in long waiting lists and added burdens for children and families who are in need of help.

Similar capacity problems have been described in other countries [64, 65]. International research has shown a gap between those who need and those who receive mental health services that ranges from 32% to 78% [66]. As the gap between need and capacity widens, cost-effective procedures are of increasing importance, but only way to achieve effective treatment is through accurate assessment and diagnosis. Indeed, misdiagnoses of mental health problems may result in delayed treatment and worsen the course of the disorder [67]. If less time is spent on the assignment of diagnoses and severity of mental health problems, referrals to appropriate

treatment can be more rapid. This could potentially increase treatment capacity, and decrease the long waiting lists in CAMHS.

More effective use of clinical resources through screening for mental health problems

Screening instruments, like the Strengths and Difficulties Questionnaire (SDQ) and the Achenbach System of Empirically Based Assessment are potentially cost-effective instruments because they aim to halt and reverse the progression of mental health problems by detecting them more effectively. Screening in multiple settings with input from multiple informants is preferable, due to variability in behavior across settings and time [68]. Of particular importance to all screening instruments are the psychometric properties such as predictive validity, i.e., sensitivity, specificity, PPV, NPV, positive likelihood ratio (LHR^+), negative likelihood ratio (LHR^-) and diagnostic odds ratio (OR^D).

Sensitivity and specificity are one way of quantifying the diagnostic accuracy of a test [69, 70]. Sensitivity is the ability of the screening instrument to generate a true positive result for someone with the diagnostic category of interest. Specificity is the ability of the instrument to generate a true negative result for someone without the diagnostic category of interest [71]. Sensitivity and specificity are important to clinicians because these measures indicate how many people with disorders the screening instrument can correctly identify.

Sensitivity and specificity are also important from a population perspective in order to determine diagnostic accuracy, but for patients and their clinicians PPV, NPV, LHR^+, LHR^- and OR^D may be more informative, as they show the probability of a disorder, given a positive or negative screening result [72]. PPV and NPV refer to the probability that a positive or negative screening result reflects the correct diagnosis [71], and these values vary according to the

prevalence of a disorder in a given population [73]. For example, the PPV for a disorder with low prevalence can be low even if the sensitivity and specificity are high.

LHRs are ratios of probabilities, and are used to summarize diagnostic accuracy on the basis of sensitivity and specificity [74]. The LHR provides information on how a positive or negative screening result changes a person's likelihood to have a certain diagnosis. A single measure that summarizes the discriminative ability of a screening instrument is the OR^D. The OR^D is relatively independent of changes in both spectrum and prevalence, and is therefore a robust measure for dichotomized results. For clinical purposes 'acceptable' accuracy will vary depending on the aim (i.e., to confirm the absence or presence of a disorder), and on the possible consequences for the patient.

Setting cut-off points is difficult, as the costs and benefits of misclassifications must be balanced. Information about the natural, untreated history of the disorder, information about the adverse and beneficial consequences of treatment, information about the psychometric properties of the screening instrument, and information about subgroups with interacting risk factors in a specific population may guide such decisions [75]. In early-phase screening, high sensitivity may be of particular importance, while the importance of PPV, NPV, LHR^+, LHR^- and OR^D may be increased with successive screening rounds. Moreover, whereas false positives could result in unnecessary evaluations and thereby drain CAMHS resources, false negatives may delay the identification of mental health problems, which is strongly associated with parents' dissatisfaction with the diagnostic process [17].

More effective use of clinical resources through telepsychiatric assignments

Telepsychiatric diagnostic assignments and web-based, online instruments may also be cost-effective. Research has shown that:

> "Telepsychiatry can serve a broad spectrum of children and adolescents across demographic, socioeconomic, and payor status with a range of psychiatric diagnoses. Furthermore, these demographics, payor, and clinical characteristics of youth referred for telepsychiatry are very similar to youth evaluated in a "face-to-face" child and adolescent psychiatry outpatient clinic…. This similarity of diagnoses suggests that telepsychiatry provides adequate technical resolution and interpersonal rapport to detect the psychopathology of children and adolescents referred for outpatient care" (p. 283-284) [76].

With telepsychiatry patients benefit from reduced travel time, less time lost from school or work, shorter waiting periods and more availability of expertise [77]. Telepsychiatry may also hasten the implementation of effective treatment [78]. Good to excellent diagnostic agreement, as well as high provider and patient satisfaction, has been reported for telepsychiatric solutions [79, 80]. A Norwegian study that investigated geographic, climatic and travel-related factors found that access to, and use of telepsychiatry were widespread in Northern Norway due to long distances, climatic difficulties and low access to local mental health professionals [81]. However, a recent review of research on telepsychiatric assessment of child and adolescent behavioral disorders concluded that there are significant weaknesses associated with telepsychiatric studies, namely considering diagnostic assignments in conditions where there is a lack of competency or capacity for local follow-up, but also considering sample sizes that are not large enough to detect clinically meaningful differences between routine clinical assessments and telepsychiatric assessments [82]. However, the authors concluded that, "There are currently no findings suggesting that telepsychiatric assessments are biased toward recognizing certain disorders over others, or that telepsychiatric assessments are not comparable to in-person assessments" (p. 715) [82].

Most studies of telepsychiatric diagnostic assignments have involved two-way videoconferencing. Although online assignments share most of the benefits of two-way videoconferencing in telepsychiatry (e.g., reduced travel time, less time lost from school or work, shorter waiting periods and availability of expertise), only limited research has been conducted. The only study available to my knowledge examined agreement between online DAWBA-generated diagnoses and clinical diagnoses, as well as the effects of online assessment of child and adolescent psychopathology on clinical decision making [83]. Acceptable agreement was found between the DAWBA-generated diagnoses and clinical diagnoses ($\kappa = 0.26$-0.31), and a pronounced effect was found on clinical diagnoses of emotional disorders when the clinicians were informed about the DAWBA-generated diagnosis. More research has been conducted on web-based interventions. Two recent reviews concluded that web-based interventions, especially those with therapist support, were effective for several mental health problems (e.g., depression, anxiety, traumatic stress). Effects of web-based interventions were comparable to those of clinician-delivered interventions, and advantages over clinician-administered treatments included cost-effectiveness and accessibility. However, few web-based interventions were aimed at children and adolescents despite high internet usage among children, adolescents and their parents [84, 85]. In sum, little knowledge exists on online diagnostic assignments in children, but existing studies on closely-related fields have reported promising results.

Agreement between research and clinical diagnostic assignments

Use of DAWBA information to assign psychiatric diagnoses, collected either by lay interviewers or online, has been reported to be reliable and is common in epidemiological research [61, 86-88]. However little is known about the reliability of such methods in clinical samples. If good diagnostic accuracy can be established through web-based procedures, there is a

huge potential for saving time and clinical resources in the assessment phase, and thereby improve treatment accessibility. High agreement between clinicians using web-based information for assignment of both diagnoses and severity of mental health problems is a first step towards valid procedures. A next step might be high agreement between diagnoses assigned by clinicians using web-based information and routine clinical assignments. A study by Jensen-Doss and Weisz [89] suggested that agreement between clinician- and research-generated diagnoses may predict a successful treatment process and outcome, and research that examines variables that lead to agreement and differences between these diagnoses is called for.

Possible factors that can contribute to disagreement between research-generated diagnoses, and clinical diagnoses, here represented by online clinical assignment, and routine clinical assignment, respectively, have been suggested by Angold [90]. On the part of the clinical practitioner, they included decisions based on familiarity of different diagnoses, selectively collected information and making diagnostic decisions before all information is available. Haine et al [91] concluded that, "Clinicians may assign diagnoses on broad impressions of the domain in which a youth's problems fall, rather than on whether that youth meets full criteria for diagnoses within the particular domain." (p. 724). In addition, organizational and other non-clinical factors (i.e., socially acceptable diagnoses, social service regulations, insurance regulations) may influence diagnostic assignment in routine clinical practice, but not in a research setting.

General research questions

The main aims of the work undertaken within the framework of this dissertation were to investigate the standardized assessment instruments used in routine clinical practice to assign diagnoses and severity of mental health problems. Both the reliability and the validity of some instruments have been examined and one main focus was the clinical usefulness of these instruments and their potential for more effective use of limited clinical resources. The main research questions addressed in the three papers resulting from this dissertation are outlined below:

The purpose of the first paper was to examine the agreement between diagnoses and severity ratings assigned by clinical specialists who were trained CAMHS researchers based only on DAWBA information collected online.

The aim of the second paper was to examine whether the application of specific scoring algorithms for the SDQ could be useful in screening for mental health disorders among children and adolescents in the CAMHS North Study

The purpose of the third paper was to examine the agreement between diagnoses and severity assigned by clinical specialists trained as CAMHS researchers, based only on DAWBA information collected online, and the routine clinical assignments by CAMHS clinicians. We also wanted to examine factors contributing to agreement or disagreement on these diagnoses.

Methods of papers 1, 2 and 3

Participants

All individuals aged 5 to 18 years, referred for diagnostic assessment to either the Child and Adolescent Mental Health Outpatient Clinic at the University Hospital of Northern Norway, or to the Alta Child and Adolescent Mental Health Outpatient Service at the Finnmark Hospital Trust, by either a general practitioner or child social welfare authorities, during the period September 2006 to December 2008 were invited by mail to participate (N = 1,032) in the CAMHS North Study. Written consent was collected for a total of 286 patients (28%). A significantly higher number of oral consents were registered and almost no refusals to participate were collected. The participants in the CAMHS North Study included 155 boys (54%) and 131 girls (46%) with a mean age of 11.11 years (standard deviation = 3.35, range = 5-18 years). There were a total of 128 (45%) children (5-10 years old, 65% boys) and 158 (55%) adolescents (11-18 years old, 46% boys). Norwegian national statistics for CAMHS [20] shows a similar distribution for sex and age, with more boys (57%) than girls, and more adolescents (60% 13 years old or above) than children. Parents of participating patients provided information on their ethnicity (85% non-immigrant Norwegian, 3% Sami people, 4% immigrants from Europe), parental status (47% both biological parents, 27% one biological parent, 13% one biological parent and his/her new partner, 4% foster care), household income (56% double income, 26% single income), socioeconomic stress (72% none/minor, 14% major), stress associated with work and work pressure (63% none/minor, 23% major), and stress associated with physical and mental health (71% none/minor, 15% major), which was recorded in the DAWBA background module (missing data for 8-18%).

Procedure for online clinical assignment

Parents, teachers and children above the age of 11 years completed the relevant version of the DAWBA using a web-based interface that they accessed from home or school after receiving a request with the unique web link for that child's case. Written information about how to log on, as well as contact information in case of problems, was distributed along with a unique web ID and password. For participants younger than 16 years of age, requests were distributed by mail to the parents, who in turn distributed the requests to their children (if aged 11-15 years) and the teachers. For the participants 16 years of age or older, requests to both parents and teachers were distributed by the participants themselves.

Four experienced clinicians independently assessed the participants of the CAMHS North Study. Of the four rating clinicians, three were clinical specialists in neuropsychology with a minimum of 9 years of experience in the field, and one was a specialist in child and adolescent psychiatry with 15 years of experience in the field. All clinicians completed the online training for the DAWBA [92]. They also completed a 1-day training session on the categories of severity in the Children's Global Assessment Scale (C-GAS) and Health of the Nation Outcome Scale for Children and Adolescents (HoNOSCA), including scoring of vignettes [93, 94]. In addition, all four clinicians participated in two separate 2-day training sessions in preparation for this study, including diagnostic assessment and severity ratings of clinical cases. The clinician who led the 2-day training sessions was trained by Robert Goodman, who developed the DAWBA interview.

Each clinician individually diagnosed the participants according to the ICD-10 diagnostic criteria for research [95]. The assessment was based on information collected from parents, teachers and/or self-report through the DAWBA, without face-to-face contact with the parents, teachers or participants themselves. The available information was identical for all four clinicians. To ensure enough cases for agreement analysis, the diagnoses were categorized as

emotional disorders (diagnoses related to separation anxiety, specific phobias, social phobia, panic attacks and agoraphobia, post-traumatic stress disorder, generalized anxiety, compulsions and obsession, depression, and deliberate self-harm), attention deficit hyperactivity disorder (ADHD)/hyperkinetic disorders (diagnoses related to attention and activity), conduct disorders related to awkward and troublesome behavior), and other disorders (diagnoses related to developmental disorders, eating difficulties, and less common problems). Comorbidity was documented when diagnoses from at least two categories were assigned, without taking the exclusion rules of the ICD-10 into consideration. The clinicians also assigned clinical severity according to the C-GAS and HoNOSCA.

Procedure for routine clinical assignment

All patients receiving care from the CAMHS, according to Norwegian guidelines for CAMHS [13], undergo clinical assessment to assign diagnoses and severity. The assignment of diagnoses was based on multi-professional consensus discussions with at least one attending psychiatrist or clinical specialist in psychology, where all available information from different sources (i.e., clinical history, clinical diagnostic interviews, cognitive assessment), including the DAWBA, were integrated into an assignment of diagnoses according to the clinical description and diagnostic guidelines in the ICD-10 [96]. It is uncertain to what degree the available DAWBA information was used by the clinicians. There are no formal or clearly replicable procedures for routine clinical assignments of diagnoses. Clinical information may have been collected by clinicians with different professional backgrounds, there were no fixed multi-professional groups for the consensus discussions, and the timing of the assignment of diagnoses differed over the course of 'patient status'. This means that some diagnoses may have been assigned after just one face-to-face meeting with the patient, while other diagnoses were assigned

at the end of a treatment period lasting several years. The assignment of severity was based on information collected in a single 1-hour face-to-face referral meeting, taking place a maximum of 10 days after referral, and at which time no DAWBA information was available to the clinicians.

Measures

Information contained in the DAWBA was used by the clinicians to assign ICD-10 diagnoses and C-GAS and HoNOSCA severity ratings of mental health problems.

Development and Well-Being Assessment

The DAWBA is a package of measures of child and adolescent psychopathology for administration to multiple informants. It is designed to generate common child psychiatric diagnoses according to the ICD-10 and DSM-IV, without neglecting severe, but less common diagnoses. The Norwegian web-based version that was used in the CAMHS North Study contains modules for diagnoses related to separation anxiety, specific phobias, social phobia, panic attacks and agoraphobia, post-traumatic stress disorder, generalized anxiety, compulsions and obsession, depression, deliberate self-harm, attention and activity, awkward and troublesome behavior, developmental disorders, eating difficulties, and less common problems, as well as modules for background information and strengths. For each module there are both closed questions with fixed response categories and open-ended questions where the informant is asked to give detailed descriptions in his/her own words in text-boxes. Each module has initial screening questions with skip rules, and if problems are reported informants are also asked about their functional impact. Three different versions are available: 1) a detailed psychiatric interview for parents of approximately 50 minutes in length, 2) a youth interview of approximately 30 minutes and 3) a briefer questionnaire for teachers of approximately 10 minutes. The information from all informants is presented to the clinician in a separate program, where all closed questions are used

to generate predictions of likelihood for a diagnosis [92]. The predictions can be used as rough prevalence estimates for research purposes [97], but mostly as a convenient starting point for clinicians evaluating all information, including the open-ended questions, in order to determine the correct diagnoses for the child. The DAWBA has shown good discriminative properties both between population-based and clinical samples, and between different diagnoses [98]. Both in Norway and the UK, the DAWBA has been shown to generate realistic estimates of prevalence for psychiatric illness, as well as to have a high predictive validity when used in public health services [61, 87]. Good to excellent inter-rater reliability has been reported in both British and Norwegian studies, with $\kappa = 0.86$ to 0.91 for 'any disorder' $\kappa = 0.57$ to 0.93 for emotional disorders, and $\kappa = 0.93$ to 1.0 for ADHD/hyperkinetic or conduct disorders [86, 99]. Good to excellent agreement has also been reported between routine clinical diagnostic assignments and those based solely on the DAWBA, with κ ranging from 0.57 to 0.76 [39, 100].

Strength and Difficulties Questionnaire

The SDQ is a screening instrument embedded in the DAWBA that covers problems and resources relevant to the mental health and behavior of children and adolescents aged 4 to 16 years [101]. There are three different versions: the parent version and teacher version rate behavior for all ages; a self-reported version is used only among adolescents aged 11 to 16 years. The SDQ contains 25 items, covering five areas of clinical interest: hyperactivity/inattention (e.g., 'restless, overactive, cannot stay still for long'), emotional symptoms (e.g., 'many worries, often seems worried'), conduct problems (e.g., 'often has temper tantrums or hot temper'), peer relation problems (e.g., 'picked on or bullied by other children') and prosocial behavior (e.g., 'kind to younger children'). The extended version of the SDQ also covers severity of difficulties, chronicity, overall distress, social and scholastic impairment, and burden to others (e.g., 'how

long have these difficulties been present', 'do the difficulties upset or distress your child', 'do the difficulties interfere with your child's everyday life in the following areas') [102].

Based on both symptoms and the corresponding impact reported by parents, teachers and self-report, predictive algorithms have been developed for a broad category, 'any disorder', as well as for three subcategories: conduct disorders, hyperactivity disorders, and emotional disorders. These algorithms, which are based on established British norms/cut-offs, have been tested in several cultures. They are described in detail by Goodman et al [103] and syntaxes are available online (www.sdqinfo.org), where normative data from different countries can be found. Country, gender and age affect the exact proportion of classifications, but these algorithms will classify approximately 80% of a population-based sample as 'unlikely' to have a psychiatric disorder, approximately 10% as 'possibly', and another 10% as 'probably' having a psychiatric disorder.

Children's Global Assessment Scale

The C-GAS was used to rate severity of mental health problems. It is frequently used for this purpose and has several areas of application, such as to quantify impairment levels, as an outcome measure, or as an indicator of prognosis [104, 105]. The C-GAS is a single-factor measure of the overall severity of psychiatric disturbance, with a summary score ranging from 1 to 100 that allows for a clinically meaningful index of global psychopathology. Green et al [106] found that when used in clinical practice, C-GAS measures functional strengths. Several studies have revealed good inter-rater reliability, especially among raters that have experience with C-GAS [106-108].

Health of the Nation Outcome Scale for Children and Adolescents

The HoNOSCA was also used as a measure of severity of mental health problems in the studies carried out for this dissertation. The HoNOSCA is a broad measure of behavioral, symptomatic, social, and impairment domains in children and adolescents. A total of 13 clinical

features were rated by clinicians on a five-point severity scale and added into a summary score, ranging from 0 (no problems) to 52 (severe problems in relation to all clinical features). Several studies have found good inter-rater reliability for the total score, as well as for the majority of individual items [109-112].

Statistical analyses

All statistical analyses in this dissertation were performed using either STATA version 11.0 or SPSS version 16.

Screening for psychiatric diagnoses

In order to calculate the screening efficiency of the SDQ, results were dichotomized on the original probability categories in the SDQ scoring algorithm (unlikely, possible, and probable). In a first instance calculations were made where the categories unlikely and possible were labeled 'negative' and the category probable was labeled 'positive' (hereafter referred to as 'probable' dichotomization level). In the second calculation only the category unlikely was labeled 'negative' and the categories possible and probable were labeled 'positive' (hereafter referred to as the 'possible' dichotomization level). Applying the 'probable' dichotomization level will yield a negative test result for approximately 90% of a population-based sample as having a negative test, whereas the 'possible' dichotomization level will yield a negative test result for approximately 80% of the same sample.

Sensitivity and specificity are another way of quantifying the diagnostic accuracy of a test, and so sensitivity (*sensitivity = a / (a + c),* see Table 1) and specificity (*specificity = d / (b + d)* of the SDQ was calculated. Sensitivity and specificity are important to clinicians because these measures indicate how many people with disorders the SDQ can correctly identify.

To highlight the probability of a disorder given a positive or negative screening result, PPV (*PPV = a / (a + b),* see Table 1) and NPV (*NPV = d / (c + d),* see Table 1) were calculated. To summarize diagnostic accuracy on the basis of sensitivity and specificity, LHRs (LHR^+ = *sensitivity / (1 – specificity),* LHR^- *= (1 – sensitivity) / specificity*), see Table 1) were calculated, in addition to the OR^D (LHR^+ / LHR^-, see Table 1). For clinical purposes 'acceptable' accuracy will vary depending on the aim (i.e., to confirm the absence or presence of a disorder) and depending on the possible consequences for the patient.

Table 1. Performance of a screening test

		'Gold standard'		
		Diagnosis	**No diagnosis**	**Total**
SDQ	Test positive	*a*	*b*	*a + b*
	Test negative	*c*	*d*	*c + d*
	Total	*a + c*	*b + d*	*a + b + c + d*

Note: *a = True positive, b = False positive, c = False negative, d = True negative.*

Agreement between clinicians assigning diagnoses and severity of mental health problems

For the exact proportion of cases where all four clinicians agreed on the diagnoses, raw agreement was calculated. Both precision and accuracy are important components of the inter-rater agreement of clinician-assigned diagnoses. Precision is the repeatability of the clinical assignment, or the agreement between multiple clinicians. High precision is a requirement, but not a guarantee of good accuracy, because systematic errors inherent in the instrument itself will

not be discovered [113]. To examine the agreement on diagnoses between the four clinicians, Fleiss' κ for ordinal data was calculated. Fleiss' κ measures the overall agreement for all four clinicians, without any reference to the consensus diagnoses [114].

Intra-class correlation (ICC) between clinicians was computed to assess agreement on HoNOSCA and C-GAS severity ratings. The preferred model for ICC was an alpha model for dichotomous data, and a two-way mixed type for consistency data [115, 116]. The ICC was calculated as a single-measure ICC and an average-measure ICC, where the single-measure ICC was the reliability of the ratings of one clinician, and the average-measure ICC was the reliability of the ratings of all four clinicians averaged together. The correct measure to use depends on the clinical or research situation. If the rating of only one clinician is used, the single-measure ICC is appropriate. If multiple ratings are available, it is more appropriate to use the average-measure ICC, keeping in mind that multiple ratings generally increase reliability [117].

Agreement between online and routine clinical assignment

Chi-square analyses were conducted to compare findings for clinicians and researchers. In addition, accuracy, or agreement between the online clinical assignment and the routine clinical assignment, was calculated. Accuracy usually refers to the agreement between the clinician-assigned diagnoses and an assigned 'reference', or the ability to distinguish between patients with and without the diagnoses of interest. Good accuracy means a minimum of both random and systematic errors. To examine the accuracy of the online clinical assignments, each assignment was tested against the routine clinical assignment, and Cohen's κ was calculated. As does Fleiss' κ, Cohens' κ measures inter-rater agreement, but is limited to the agreement between two raters or measures [114].

Identification of patterns of agreement and disagreement

In this paper three multinomial logistic regression was used to identify whether different patterns of agreement or disagreement between routine clinical assignments (RCA) and online clinical assignments (OCA) were associated with different predictors. Due to the lack of a real 'gold standard' for psychiatric diagnoses, simple logistic regression may not reveal the differences that are important to understanding the consequences of choosing online or routine clinical assignment. The presence and absence of diagnoses were indicated with $^{+}$ and $^{-}$, respectively. Both patient- and service-related factors (age, gender, urban or semi-rural clinic, time gap between online and routine clinical assignment, difference in C-GAS scores between online and routine clinical assignment, difference in HoNOSCA scores between online clinical and routine clinical assignment, and number of informants) were entered in a multinomial logistic regression model. Multinomial logistic regression estimated the log odds of each of the four outcomes ('agreement on presence of diagnoses' [$RCA^{+}OCA^{+}$], 'agreement on absence of diagnosis' [$RCA^{-}OCA^{-}$], 'disagreement, absent routine clinical diagnosis/present online clinical diagnoses' [$RCA^{-}OCA^{+}$], 'disagreement, present routine clinical diagnoses/absent online clinical diagnosis' [$RCA^{+}OCA^{-}$]) compared to a reference. This analysis was repeated with different references to simplify the interpretation of the results.

Guidelines for interpretation of results

Interpretations of κ values followed the guidelines suggested by Cicchetti and Sparrow [114]. Agreement in the range $\kappa = 0.75$ to 1.00 were interpreted as excellent, $\kappa = 0.60$ to 0.74 as good, $\kappa = 0.40$ to 0.59 as fair, and $\kappa < 0.40$ as poor.

The interpretations of the ICC values were done according to the guidelines suggested by Shrout [117]. Agreement in the range of 0.81 to 1.00 was interpreted as substantial, 0.61 to 0.80 as moderate, 0.41 to 0.60 as fair, 0.11 to 0.40 as slight and 0.00 to 0.10 as virtually none.

The LHR^+, the LHR^-, and the OR^D were interpreted according to the rule of thumb described in Fischer et al [118], where potentially useful tests (i.e., those that may alter clinical decisions) are usually characterized by an LHR^+ greater than 7, an LHR^- less than 0.3, or an OR^D above 20.

Ethical considerations

From an ethical point of view, research on patients is more complicated than research on healthy people, and research involving children is more demanding than research on adults. Combining these two factors to carry out research on child patients may well be one of the fields that poses the greatest demands for thorough ethical reflection. The risks of participation are minimal, but no research can be described as entirely free from risk in terms of psychological damage [119].

Written informed consent was obtained for all patients before inclusion in the CAMHS North Study. Parents gave consent for patients under 12 years of age. For patients between 12 and 16 years of age, written consent was obtained from both the parents and the patients. Patients over 16 years of age gave consent themselves according to Norwegian legislation. The Regional Committee for Medical Research Ethics (REK) and the Norwegian Social Science Data Services approved the study. Despite this, it is appropriate to reflect upon research on children in general, and in the field of child and adolescent mental health in particular. From an ethical viewpoint, it is imperative to conduct research on child and adolescent psychiatric problems. In terms of the various declarations and regulations, there is clear guidance for how such research should be designed [120], but different terminologies are used in various regulations [121, 122]. Vitiello et al writes:

> "Most of the controversy that surrounds the ethics of conducting research in children lies not so much on the general principles that regulate such research, but on the applications

of the regulations and interpretation of how concepts such as minimal risk, minor increase over minimal risk, knowledge of vital importance, and favorable risk/benefit ratio apply to the specific research project under consideration" (page 1048) [123].

It is difficult to assess the ethical aspects of one's own research from an unbiased point of view. Special interests can be compelling and it is easy to minimize any inconvenience or risk that may be borne by others [124]. Of course an approval from the REK is no guarantee that ethical pitfalls or difficult decisions not will arise during a project.

Morrow and Richards [125] argued that the greatest ethical challenge for research involving children is the difference in power and status between children and adults. There is a broad consensus that patients, and especially child patients, should be considered a vulnerable group. As such they are afforded special protection when they participate in research [122]. One way to protect children is informed consent, which is the standard in all medical research. Although research that includes children is the only area in Norway that allows vicarious consent [126], even children who are not considered competent to consent should, to the extent that it is possible, give their assent, i.e., a confirmation of agreement to participate [127]. Research that includes children, regardless of parental consent, should only be carried out if the child does not oppose it. In addition, the requirements of negligible risk/disadvantage, of benefits for the child or others with the same age-specific condition, and that similar research not be conducted on other individuals who are competent to consent, must be fulfilled [121].

A European study questioned the validity of parental consent, and as an extension asked whether consent should be an absolute requirement to include children in research [128]. Validity in relation to consent was assessed using four criteria: competence (the person giving consent is mentally competent to do so), information (sufficient information is received to give informed consent), comprehension (understanding is sufficient to make a deliberate choice) and

volunteerism (the consent is given voluntarily). Consent was considered valid when all four criteria were met, and by this measure a large majority of consents (70 %) were considered to be invalid. Yet the majority of parents reported that the consent process was valuable and that they felt involved in the decision for their child to participate. One argument against requiring informed consent in medical research is that the consent process is an unnecessary burden [129], but Mason and Allmark [128] found little support for the view that in some cases parents should be "spared" a request for consent.

A child's right to choose when it comes to participation in research, which is regulated by the Convention on the Rights of the Child, include the right to participate in research [130]. Children are important contributors to many research areas, including mental health. Research on children is necessary to understand both normal function and development of psychiatric pathology. Research is also important to develop effective prevention and treatment, and to reduce the impact of mental disorders in patients, their families and communities. During childhood and adolescence, major changes occur, both structurally and functionally. This may affect factors such as manifestation of symptoms, response to medication and other treatments, etc. Extrapolating research findings from adults to children is not always possible, and leads to a lack of understanding of children's conditions. Even more important is that such extrapolation may have consequences in the form of erroneous conclusions about treatment, causing illness and injury [131].

The inclusion of children who are also patients in medical research presents researchers with a number of difficult ethical considerations. The first is whether it is necessary to perform research on child patients, or whether the research can be accomplished in a less vulnerable group. The dissertation project deals with children with mental disabilities and showed major differences in symptoms and diagnostic criteria in children compared to older patients,

reinforcing that results from research on adult patients are not necessarily transferable to children. Furthermore, the relevant ethical review boards considered that there was minimal risk associated with participation in this project, that the project would provide "knowledge of vital importance" for the patient group, and that there was a "favorable risk/benefit ratio". All medical research in Norway needs to be approved by the REK and this independent approval was obtained before the start of this project. My tasks during the research process included regularly monitoring the ethical issues. New knowledge can change the assessment of "knowledge of vital importance", or new methods can change the "risk/benefit ratio" in the project.

Also the participants must be able to consider the ethical qualities of the project both before they consent and during the study. In my project we have, to the greatest extent possible, tried to help children and parents to make informed decisions about consent. Information was given both orally and in writing, and was explained to any and all individuals who so wished. Although the consent process can be a burden, it is one that the vast majority of parents prefer to take on, if the alternative is having others make the decision about their child's participation [128].

Research on child patients is complicated, but children have the right to participation, and knowledge of child patient populations is important. To exclude children from participating in research may have major negative consequences in for both the short and long term. An absolute requirement for research is that it meet all criteria of ethical responsibility and that this be evaluated regularly throughout the process.

Summary of papers

First paper: Inter-rater reliability for diagnoses and severity of mental health problems

Brondbo PH, Mathiassen B, Martinussen M, Heiervang E, Eriksen M, Kvernmo S. Agreement on web-based diagnoses and severity of mental health problems in Norwegian child and adolescent mental health services. Clin Pract Epidemiol Ment Health 2012; 8:16-21.

Objective

This study examined the agreement between diagnoses and severity ratings assigned by clinicians using a structured web-based interview within a child and adolescent mental health outpatient setting.

Methods

Information on 100 youths was obtained from multiple informants through a web-based DAWBA. Based on this information, four experienced clinicians independently assigned diagnoses (according to the ICD-10) and severity of mental health problems according to the HoNOSCA and the C-GAS.

Results

Raw agreement between the four clinicians was calculated for both a dichotomous level of agreement on any diagnosis versus no diagnosis and a second level for agreement on the type of clustered diagnoses. For 'any disorder' raw agreement was 75%, for emotional disorder it was 77%, for ADHD/hyperkinetic disorder 84% and for conduct disorder 84%.

Fleiss' κ was used to examine the precision of the clinician-assigned diagnoses. We found that the precision of the diagnoses was good, both for the dichotomous criterion of diagnosis/no diagnosis ($\kappa = 0.69$, 95% confidence interval [CI] 0.66-0.73) and for the different sub-types of

diagnoses: emotional disorder = 0.70 (95% CI 0.68-0.75) and ADHD/hyperkinetic disorder diagnosis $\kappa = 0.72$ (95% CI 0.68-0.76). For conduct disorder the precision was excellent, $\kappa = 0.82$ (95% CI 0.76-0.87).

The ICC of the clinician-assigned severity of mental health problems for single measures was moderate for both the total score of HoNOSCA (0.80, 95% CI 0.74-0.85), and for C-GAS (0.76, 95% CI 0.69-0.82). For average measure the ICC was substantial for both the total score of HoNOSCA (0.94, 95% CI 0.92-0.96) and for C-GAS (0.93, 95% CI 0.90-0.95).

Conclusions

Agreement was good to excellent for all diagnostic categories. Agreement on severity was moderate, but improved to substantial when the average of the ratings given by all clinicians was considered. Therefore, we conclude that experienced clinicians can assign reliable diagnoses and assess severity based on DAWBA data collected online.

Second paper: Screening for mental health disorders in clinical practice

Brondbo PH, Mathiassen B, Martinussen M, Heiervang E, Eriksen M, Moe T, Saether G, Kvernmo S. The strengths and difficulties questionnaire as a screening instrument for Norwegian child and adolescent mental health services, application of UK scoring algorithms. Child Adolesc Psychiatry and Ment Health 2011; 5:32.

Objective

The use of screening instruments can reduce waiting lists and increase treatment capacity. The aim of this study was to examine the usefulness of the SDQ with the original UK scoring algorithms, when used as a screening instrument to detect mental health disorders among patients in the Norwegian CAMHS North Study.

Methods

A total of 286 outpatients, aged 5 to 18 years, from the CAMHS North Study were assigned diagnoses based on the DAWBA. The main diagnostic groups (emotional, hyperactivity, conduct and other disorders) were then compared to the SDQ scoring algorithms using two dichotomization levels: 'possible' and 'probable' levels. Sensitivity, specificity, PPV, NPV, LHR^+, LHR^-, and OR^D were calculated.

Results

As expected, the amount of SDQ-predicted diagnoses was highest when the 'possible' dichotomization level was applied for all disorders. For the prevalence of 'any disorder', the 'possible' dichotomization level was 89%, compared to 72% for the 'probable' dichotomization level, and 66% for the DAWBA-generated diagnoses. In addition, the rates of SDQ-predicted diagnoses using the 'probable' dichotomization level were higher than the rates of DAWBA-generated diagnoses for all categories except emotional disorders.

A total of 66% of patients were assigned a psychiatric diagnosis based on the DAWBA, and of those almost one-third (21%) were assigned comorbid diagnoses. A diagnosis of emotional disorder was assigned to 34% of patients. A diagnosis of hyperactivity disorder was assigned to 18% of patients. Conduct disorder diagnoses were assigned to 31% of patients. Other disorders were assigned to 7% of the patients. The most common comorbid diagnoses were hyperactivity disorder in combination with conduct disorder (10%) and emotional disorder in combination with conduct disorder (8%). Sensitivity for the diagnostic categories included was 0.47 to 0.85 ('probable' dichotomization level) and 0.81 to 1.00 ('possible' dichotomization level). Specificity was 0.52 to 0.87 ('probable' dichotomization level) and 0.24 to 0.58 ('possible' dichotomization level).

Also the discriminative ability varied due to the different levels of dichotomization. When the 'probable' dichotomization level was applied, none of the LHR^+ results (1.78-3.91) were in the interval for potentially useful tests. The categories hyperactive disorders, conduct disorders, and 'any disorder' were all in the LHR^- interval for potentially useful tests (0.23-0.29). None of the OR^D results were in the interval for potentially useful tests as indicated by the guidelines provided by Fischer et al [118]. After applying the 'possible' dichotomization level, none of the LHR^+ results (1.25-2.30) were in the interval for potentially useful tests. The categories hyperactive disorders, conduct disorders, and 'any disorder' were all in the LHR^- interval for potentially useful tests (0.00-0.18). Likewise, the OR^D results for hyperactive disorders and conduct disorders were in the interval for potentially useful tests (39.26-∞).

Conclusions

The usefulness of the SDQ UK-based scoring algorithms in detecting mental health disorders among patients in the CAMHS North Study is only partly supported in the present study. They seem best suited to identify children and adolescents who do not require further psychiatric evaluation, although this is also problematic from a clinical point of view.

Third paper: Comparing online and routine clinical assignments

Brondbo PH, Mathiassen B, Martinussen M, Handegard BH, Kvernmo S. Agreement on diagnoses and severity of mental health problems between a research and a naturalistic clinical setting. J Telemed Telecare submitted.

Objective

The purpose of this study was to examine the agreement between diagnoses based on DAWBA information collected online, and routine diagnostic assignment by CAMHS clinicians.

Factors contributing to agreement or disagreement on diagnoses between the online and routine clinical assignment were also examined.

Methods

Routine clinical assignment of diagnoses for 286 patients from the CAMHS North Study were compared to those from an online clinical assignment based on information from the DAWBA. Chi-square analysis, kappa statistics and multinomial logistic regression were performed.

Results

Raw agreement on the different diagnostic categories varied between 74% and 90%, resulting in κ values in the fair range (0.41-0.49). The final model for emotional disorder had a chi-square of 53.05 (df = 21, $p < 0.001$) and a pseudo R-square (Nagelkerke) of 0.22. Age ($\chi^2 = 20.24$, $p < 0.001$), and gender ($\chi^2 = 10.22$, $p < 0.05$) were factors that significantly contributed to the explanation of different patterns of agreement and disagreement. The final model for ADHD/hyperkinetic/conduct disorder had a chi-square of 58.32 (df = 21, $p < 0.001$) and a pseudo R-square (Nagelkerke) of 0.24. Age ($\chi^2 = 21.82$, $p < .001$), and number of informants ($\chi^2 = 13.34$, $p < 0.01$) were factors that significantly contributed to the explanation of different patterns of agreement and disagreement. Time between online and routine clinical assignments, difference in HoNOSCA score between online and routine clinical assignment, difference in C-GAS score between online and routine clinical assignment, and urban or semi-rural clinic were not significant factors. The results of the multinomial logistic regression indicated different predictors of agreement and disagreement on emotional and ADHD/hyperkinetic/conduct disorders

Disagreement with $RCA^{-}OCA^{+}$ as the reference

When $RCA^{-}OCA^{-}$ was compared to $RCA^{-}OCA^{+}$ for emotional diagnoses, the only significant factor was the difference in scores between online clinical and routine clinical HoNOSCA assignment. When the difference in scores increased, the odds for $RCA^{-}OCA^{-}$ decreased. For ADHD/hyperkinetic/conduct disorders the only significant factor was number of informants. When the number of informants increased the odds for $RCA^{-}OCA^{-}$ decreased.

Disagreement with $RCA^{+}OCA^{-}$ as the reference

For the comparison between $RCA^{-}OCA^{-}$ and $RCA^{+}OCA^{-}$, and for the comparison between $RCA^{-}OCA^{+}$ and $RCA^{+}OCA^{-}$ no factor was significant for emotional disorders. For ADHD/hyperkinetic/conduct disorders age, number of informants and type of clinic were significant factors when comparing $RCA^{-}OCA^{-}$ and $RCA^{+}OCA^{-}$. When age increased the odds for $RCA^{-}OCA^{-}$ increased. When the number of informants increased the odds for $RCA^{-}OCA^{-}$ decreased, and when the assignment was made by the urban clinic the odds for $RCA^{-}OCA^{-}$ decreased.

Agreement on presence of diagnoses as reference

When $RCA^{-}OCA^{-}$ was compared to $RCA^{+}OCA^{+}$ for emotional disorders, age and gender were significant factors. When age increased the odds for $RCA^{-}OCA^{-}$ decreased. For gender, the odds for $RCA^{-}OCA^{-}$ increased for 'males' compared to 'females'. For ADHD/hyperkinetic/conduct disorders age and number of informants were significant factors. Increased age increased the odds for $RCA^{-}OCA^{-}$ and an increased number of informants decreased the odds for $RCA^{-}OCA^{-}$. When $RCA^{+}OCA^{-}$ was compared to $RCA^{+}OCA^{+}$, age was the only significant factor for both emotional and ADHD/hyperkinetic/conduct disorders. When age increased the odds for $RCA^{+}OCA^{-}$ decreased for emotional disorders, but for ADHD/hyperkinetic/conduct diagnoses the odds for $RCA^{+}OCA^{-}$ increased. Finally when $RCA^{-}OCA^{+}$ was compared to $RCA^{+}OCA^{+}$ age

was the only significant factor for both emotional and ADHD/hyperkinetic/conduct disorders. When age increased the odds for $RCA^{-}OCA^{+}$ decreased for emotional disorders, but for ADHD/hyperkinetic/conduct disorders the odds for $RCA^{-}OCA^{+}$ increased.

Conclusions

Agreement on mental health diagnoses can be fair when online clinical assignments and routine clinical assignments are compared, and may be sufficient to replace routine clinical assignment of diagnoses with an online clinical assignment, thereby saving time and resources. Age, gender and number of informants contributed to agreement and disagreement on diagnoses. The changes in odds were small in magnitude and the factors probably do not consistently contribute to the understanding of agreement or disagreement in any clinically meaningful way.

Discussion

The major aim of this dissertation was to investigate some of the standardized assessment instruments used in routine clinical procedures for the assignment of diagnoses and severity of mental health problems. Both the reliability and the validity of some instruments were examined, with the main focus being the clinical usefulness of these instruments and their potential to make more effective use of limited clinical resources. The results will be discussed, but some methodological issues should first be noted.

Methodological considerations

Methodological problems occur in all projects. Reflecting on the research process and any associated weaknesses is a way to improve the quality of the material and the conclusions drawn. Research in the field of mental health care is dominated by small samples and is thus inherently under-powered [132, 133].

Selection bias

It is relevant to discuss self-selection bias when studying a group that can decide whether or not they want to participate. In practice, this applies to all forms of research that require consent, as the same differences that lead one group to participate and another not to are likely extend to other areas [134]. Selection bias is debatable on many levels. For example, only two out of 13 invited outpatient clinics contributed to the data collection for the present dissertation. This type of self-selection at a cluster level is problematic in terms of representativeness, but also affects sample size. Clinic participation may have been an effect of motivation and sense of ownership of the CAMHS North Study. Indeed, one of the participating clinics was the host of the pilot study, and the other was the host of the main study.

There is also self-selection at the individual level, as each patient must consent to participate. We did not have permission to perform dropout analyses, therefore it was difficult to determine whether our sample was representative, or if the 268 participating patients represented a subgroup with special characteristics. To examine the possible extent of the selection bias, we compared our sample to public register data for Norwegian psychiatric outpatient clinics in relation to gender and age. Furthermore, we compared the sample to national studies in relation to gender, age, diagnosis and clinician-rated global functioning [61]. The comparisons suggested that our sample was relatively similar to a "normal" Norwegian CAMHS outpatient population. However, this is not a guarantee that no systematic biases exist in the sample, as the sample may differ on other variables not examined. However, studies of self-selection and non-response bias suggest that data on health, personality and lifestyle are relatively unbiased even with moderate response rates, and that self-selection has little impact on prevalence estimates [135, 136]. In addition, for studies that focus on agreement, representativeness may be less important than a certain degree of variation.

Sample size

The possibilities for what can be investigated are considerably limited when sample sizes are small. Most of the research questions posed in the present dissertation require a relatively large sample, and can be more interesting and answered with more nuance if the sample is large enough to examine subgroups. Based on the known prevalence of various mental disorders, possible research questions in the various research projects and the expected participation rate of the patients, ethical permission was sought to collect data from 600 CAMHS patients in the period from September 2006 to December 2008. As the final sample size was 286, one may ask if that is a sufficient sample.

There is no simple answer to this question. When it comes to, for example, the presence of 'any disorder', the prevalence in Norwegian CAMHS outpatient clinics is between 50% and 90%. For gross categories such as emotional disorders, ADHD/hyperkinetic disorders and conduct disorders, the prevalence is just below 20% to just above 30% [137]. For specific diagnoses the prevalence decreases slightly for some frequently-occurring disorders such as ADHD and depression. For other rarer disorders, prevalence can be below 1% in an outpatient sample. The rarest disorders are therefore difficult to capture to a sufficient degree, even with very large samples. However, the vast majority of mental disorders with a prevalence of between 5% and 10% in an outpatient sample is both possible and desirable to capture to a sufficiently large extent with a well planned and executed study. A sample of 286 patients can well capture the broad categories and the most frequently occurring disorders. Unfortunately, it is not sufficient to simply estimate the prevalence of rarer mental illnesses, or the vast majority of mental disorders with a prevalence of between 5% and 10% in an outpatient sample. If the present study sample were doubled, more differentiated and interesting analyses may have been possible. The limitations of this study therefore include the lack of statistical power to detect factors with real, but small effects. In addition, we cannot rule out the possibility that other factors that were not investigated may contribute to agreement or disagreement in a significant way.

Limited knowledge about routine clinical procedure

Another limitation that also should be noted is the lack of knowledge about the exact procedures of the CAMHS North routine clinical assignment of diagnoses and severity. On the other hand, routine clinical assignment can be characterized by its lack of exact procedures. As long as the gain in validity from structured procedures is uncertain [42], we may have to accept a certain degree of clinical freedom. Indeed, highly structured diagnostic procedures may cause a

major loss of clinically significant findings [53]. All instruments examined in this dissertation were already implemented in the routine clinical practice of the clinicians involved. New, potentially more effective ways to use these instruments were evaluated against the routine clinical assignments. The strength of the procedure lies in its ecological validity, as the diagnostic procedure is quite similar to the routine clinical practice in Norwegian CAMHS.

Discussion of the main findings in the first paper

Reliability was the focus of the first study that examined the agreement between CAMHS researchers who were also trained as clinical specialists, when assigning diagnoses and severity of mental health problems based only on DAWBA information collected online. Our results indicated that agreement on mental health diagnoses can be good to excellent when assignment is aided by the DAWBA, and are consistent with the findings of other studies in which diagnostic agreement in mental health populations was examined [35, 38]. Despite differences in population and clinical setting, our results strengthen the claim that, when aided by structured or semi-structured instruments, agreement on mental health diagnoses can be good to excellent, even when information is collected online. Good to excellent diagnostic agreement has been previously reported for diagnoses assigned via videoconferencing [35]. Our results suggested that an online procedure for collecting information can also be sufficient for reliable diagnostic assignments.

The second aim of this study was to examine agreement between clinicians assigning severity of mental health problems, as measured with C-GAS and HoNOSCA. The use of the DAWBA as the source of information, instead of written vignettes as most other studies have used, increases the complexity and amount of information available, and thereby lessens the focus on themes that are directly relevant when rating by HoNOSCA and C-GAS. These differences improved the

ecological validity of our results, which showed that agreement on severity ratings based on DAWBA information collected online can be fair to moderate for a single clinician, and moderate to substantial when an averaged rating from multiple clinicians is used. Even with strengthened ecological validity, our results are on par with the HoNOSCA ICC, and better than the C-GAS ICC obtained by Hanssen-Bauer et al [93]. Our results were also comparable with C-GAS ICC when chart information was assigned by untrained health-care professionals, but worse than the ICC for expert raters. We believe that this phenomenon may be explained by the increased complexity and amount of information available in our study, as well as by the diminished focus on questions that directly affect HoNOSCA and C-GAS scores. A way to enhance the reliability of both HoNOSCA and C-GAS ratings is to let multiple clinicians rate the same patient. Our single-measure ICC was moderate for both C-GAS and HoNOSCA, but the average-measure ICCs were substantial for both instruments. It is noteworthy that, by using multiple clinicians, we compensated for the complexity of the DAWBA information and showed an ICC on par with the expert group of Lundh et al [138], who rated less complex vignettes.

Discussion of the main findings in the second paper

Clinical usefulness and potential for effectiveness were the focuses of the second study. The aim was to examine the application of specific scoring algorithms for the SDQ, as proposed by earlier UK findings, when used as a screening test to detect mental health disorders among children and adolescents. Overall, our results were comparable to other studies of sensitivity and specificity of the SDQ [64]. One exception was the sensitivity to detect emotional disorders, which was considerably lower than earlier findings from the UK [103]. This difference may be an effect of Norwegian parents' and teachers' 'blind spot', or 'normalizing' view for emotional difficulties, which was also reported by Heiervang et al [99]. It is also generally accepted that

parents are insensitive to children's emotional symptoms, and that adolescents' reports of emotional problems are more valid than their parents' and teachers' reports [68, 139, 140]. This knowledge may have affected the assignments of the diagnosing clinicians in our study, and resulted in lower sensitivity for emotional disorders. Another exception was the specificity for conduct disorders which was substantially higher than in the British sample. This may be due to cultural differences between the countries, in that the degree of reporting problems in the UK may be higher, whereas Norwegian parents and teachers tend to report fewer problems. In contrast to emotional disorders, the lower SDQ scores for conduct problems seems to reflect a real and substantial lower prevalence of conduct disorders in Norway compared to the UK [99].

Overall our sensitivity and specificity results strengthen the earlier reported usefulness of the SDQ as a screening instrument for mental health problems when used in epidemiological research. Regarding clinical use, despite differences in culture and language, the scoring algorithms worked equally well in the Norwegian CAMHS North Study as in English, Bangladeshi, and Australian clinics. With the most common cut-off at approximately 90%, the SDQ will correctly identify four out of five children with psychiatric diagnoses, except for emotional disorders, and will also correctly identify most children without diagnoses, except for 'any disorder', but unfortunately, many classifications will be either false positives or false negatives. Choice of cut-offs may depend on the relative importance of false positives and false negatives, respectively. For research purposes both scenarios are sufficient, but not for clinical purposes, for which the high rates of false positives are not acceptable.

Sensitivity and specificity are important from a population perspective, but for patients and their clinicians PPV, NPV, LHR^+, LHR^- and OR^D may be more informative, as they show the probability of a disorder, given a positive or negative screening result. Our results by diagnostic category, showed a high NPV and lower PPV, which were very similar to the results reported by

Goodman and colleagues [141]. This indicates that the SDQ functions considerably better as an instrument to rule out, rather than to confirm, possible psychiatric diagnoses. The pattern may be even stronger when mental health problems are combined with chronic physical illness [142].

To our knowledge, $LHR^{+/-}$ and OR^{D} have not been reported in previous studies. Our results showed that when using the most common dichotomization ('probable' level) at approximately 90%, none of the diagnostic categories are in the OR^{D} interval for potentially useful tests. However hyperactivity disorders, conduct disorders, and 'any disorders' are in the LHR^{-} interval for potentially useful tests. For a patient with a negative screening result this is good news, because it means that this result is almost certainly correct. However, for a clinician, and for patients with positive screening results, it is also important that the PPV and LHR^{+} are high in order to reduce both economic and emotional costs associated with unnecessary further evaluations of patients that are not afflicted with the disorder of interest.

Discussion of the main findings in the third paper

Clinical usefulness, validity and potential for cost-effectiveness were also the focuses of the third study. The aim was to examine the agreement between clinical researchers assigning diagnoses and severity based only on DAWBA information collected online, and the routine diagnostic assignments by CAMHS clinicians. We also wanted to examine the factors contributing to agreement or disagreement on diagnoses between the online clinical assignment and the routine clinical assignment.

Our results showed that agreement on mental health diagnoses was in the fair range when online clinical assignments and routine clinical assignments were compared. This is consistent with the findings of other studies where diagnostic agreement in mental health populations was examined [26]. Despite differences in population and clinical settings, our results strengthen the

claim that the gap between research-derived knowledge and routine clinical practice in CAMHS is a challenge that has to be dealt with. Our results are considerably weaker than those reported for the evaluation of ADHD in clinically referred youth based on DAWBA information [39]. This may be explained by the fact that both the clinical raters and the DAWBA rater in the aforementioned study were formally trained in the use of the DAWBA, while our study compared clinical researchers formally trained in the DAWBA with clinicians without such specialized training.

The second aim of this study was to examine the factors contributing to agreement or disagreement on diagnoses between the online clinical assignment and the routine clinical assignment. Our results did not indicate that organizational or service-related factors contributed to agreement or disagreement. One could however ask if the relatively high numbers of disagreements on ADHD/hyperkinetic and conduct disorders are examples of non-clinical factors such as social acceptability and accessibility to medication. In this case, the status of ADHD/hyperkinetic disorders as neuropsychiatric conditions, and the possibility to receive pharmacological treatment if an ADHD/hyperkinetic diagnosis is assigned may alter the clinical decision when the clinician is in doubt.

Another possible explanation of disagreement is the separate ICD-10 diagnostic classifications for research and for clinical usage, respectively [95, 96]. Clinical assignments involve individualized diagnostic formulations, including considerations of treatment and prognosis, while research assignments can adopt a more strict approach, aiming at high specificity of diagnoses [53]. In clinical practice, one should expect most people with mental health problems to receive a diagnosis, but many children show considerable psychosocial impairment without fulfilling the criteria for any specific diagnosis [143]. This may encourage clinicians to assign diagnoses based on broad impressions rather than on whether the patient

meets the full criteria for a diagnosis, as Haine et al [91] stated. Our results did not indicate differences in frequencies between online clinical and routine clinical diagnoses. However, for ADHD/hyperkinetic/conduct disorders the odds for agreement on 'no diagnosis' decreased when the number of informants increased. The pattern was the same for all three comparisons and may indicate that both clinicians and researchers consider the criteria of pervasiveness in ADHD/hyperkinetic diagnoses important. For routine clinical assignment additional information from other sources, in cases where there was a single online informant, may partly explain the higher frequency of ADHD/hyperkinetic diagnoses, whereas for online clinical assignment the lack of evidence for cross-situationality required for an ADHD/hyperkinetic diagnosis may have favored a conduct diagnosis.

A third and important factor may be the timing of the assignment of diagnoses, even though this factor was not significant in our data analysis. Longitudinal data show a relatively low stability of outpatient psychiatric diagnoses, with significant fluctuations over time [144]. Our online clinical assignments were all based on data from the patients' initial contact with the services, while the timelines for routine clinical assignments differed. Some routine clinical diagnoses were assigned in the initial phase, some during the course of treatment, and most were assigned at the end of treatment. This may affect the amount of information available to routine clinical practitioners, but may also add information about changes and outcomes during the treatment period.

Clinical implications

There are some clinical implications of our results. The SDQ by itself is not a sufficient screening instrument for psychiatric disorders when used among referred patients. Our results showed that the SDQ could be better utilized to detect the presence of 'any disorder', rather than more specific diagnostic categories. On the contrary, the SDQ is better at ruling out the presence

of specific categories of psychiatric disorders than ruling out the actual presence of 'any disorder'. According to our results the SDQ is best used to identify those children and adolescents who do not need further psychiatric evaluation. However, in clinical practice this is problematic since children suffering from monosymptomatic disorders (e.g., tic disorders, enuresis, eating disorders) will not be identified through with screening with the SDQ.

Another clinical implication of the results is that a single experienced clinician trained in the use of the DAWBA is usually sufficient to assign reliable diagnoses and severity of mental health problems based on information collected online in the DAWBA. However, reliability could be further improved if several independent trained clinicians contribute to the assessment of the same patient. In a clinical setting this will of course be a question of resources, but using even two independent raters is likely to raise the reliability substantially.

In summary, practicing clinicians who assign their diagnoses through routine assessment and clinical researchers who assign their diagnoses through online assessment do agree on the diagnosis of three out of four patients and agreement can be interpreted as fair. The question is: Is a fair agreement reliable enough? Due to 'noise' from both the inconsistency of symptom expression by the patient and the application of diagnostic criteria by clinicians, κ values as low as 0.2 were described as acceptable, and κ values between 0.4 and 0.6 were described as a realistic goal for agreement on clinical diagnoses in a commentary to the field trials of DSM-V [145]. Consequently, our results indicate that the information given online by patients, parents and teachers may be good enough to be used by clinicians to assign mental health diagnoses to children and adolescents without routine clinical assessment.

Further research

Further research is needed to examine the factors that contribute to agreement between online and routine clinical assignment of diagnoses. The identification of characteristics of either the patient or the other informants that might enhance the risk of disagreement would be beneficial. With a future database large enough to subdivide the overall sample, subgroup-specific algorithms could be established and reported to facilitate comparisons between different clinical samples (e.g., with respect to age, gender, diagnostic categories) as well as identification of protective and/or risk factors. Also, further research is needed on the agreement of structured instruments when used for less prevalent mental health disorders, such as sub-types of anxiety, autism and psychosis.

To understand the gap between research and clinical diagnostic assignments, better understanding of the diagnostic processes in routine clinical practice is needed to empirically examine factors contributing to agreement. Also, further research is needed on the validity of psychiatric diagnoses. Although high agreement is important for validity, it does not ensure it. More research is needed into the validity of clinician-assigned diagnoses and severity ratings according to the HoNOSCA and C-GAS, when using the online DAWBA as the main source of information. Hyman [146] argued for the incorporation of neuroscience and genetics in a diagnostic classification system. In that case, the definition and recognition of psychiatric disorders would include etiology and pathophysiology in addition to clinical symptoms and severity. The development of a good 'gold standard' to measure and diagnose mental health problems seems the best way forward in the hunt for more cost-effective assessment methods.

Overall conclusions

In conclusion, the ability of the SDQ to detect mental health disorders among patients referred to CAMHS is not sufficient for clinical purposes. When used as a screening instrument to determine whether further evaluation is warranted in a clinical CAMHS sample, the SDQ seemed best suited to identify children and adolescents who did not require further psychiatric evaluation, although this is problematic from a clinical point of view.

Information obtained with the online DAWBA may be a sound basis on which to establish reliable clinical diagnoses and severity ratings for common mental health disorders in a clinical setting. A clinical practice that includes systematic, multiple independent assignments of diagnosis and severity is preferable due to the resulting improved reliability of the severity ratings.

Agreement on mental health diagnoses can be fair when online clinical assignments and routine clinical assignments of mental health diagnoses are compared. This may be sufficient to replace the routine clinical assignment of diagnoses with an online clinical assignment in order to save time and resources.

Compared to other medical disciplines, results of diagnostic agreement for mental health problems are on par or better [147]. When major psychiatric diagnoses were compared to medical/neurological diagnoses, the conclusion was that “there is as much objective science in psychiatry as there is in most other medical specialties, which is to say an impressive but not overwhelming amount” (22) [147].

References

1. Forskningsrådet: Evaluering av Opptrappingsplanen for psykisk helse (2001-2009). Sluttrapport - Syntese og analyse av evalueringens delprosjekter. Oslo, Norway: Divivisjon for vitenskap; 2009.

2. Hunsley J, Mash EJ: Evidence-based assessment. *Annu Rev Clin Psychol* 2007, 3:29-51.

3. Aegisdottir S, White MJ, Spengler PM, Maugherman AS, Anderson LA, Cook RS, Nichols CN, Lampropoulos GK, Walker BS, Cohen G *et al*: The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction Stefania AEgisdottir. *The Counseling Psychologist* 2006, Electronic(341-382):341-382.

4. Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C: Clinical versus mechanical prediction: a meta-analysis. *Psychol Assess* 2000, 12(1):19-30.

5. Cashel ML: Child and Adolescent Psychological Assessment: Current Clinical Practices and the Impact of Managed Care. *Professional Psychology: Research and Practice* 2002, 33(5):446-453.

6. Jenkins MM, Youngstrom EA, Youngstrom JK, Feeny NC, Findling RL: Generalizability of evidence-based assessment recommendations for pediatric bipolar disorder. *Psychol Assess* 2011, [Epub ahead of print].

7. Jenkins MM, Youngstrom EA, Washburn JJ, Youngstrom JK: Evidence-Based Strategies Improve Assessment of Pediatric Bipolar Disorder by Community Practitioners. *Professional psychology, research and practice* 2011, 42(2):121-129.

8. Drugli MB, Mørch WT: Behandling som virker. In: *Dagbladet.* 2008.

9. Norsk Psykologforening: Prinsipperklæring om evidensbasert psykologisk praksis. *Tidsskrift for Norsk Psykologforening* 2007, 44(9):1127-1128.

10. Patel V, Flisher AJ, Hetrick S, McGorry P: Mental health of young people: a global public-health challenge. *Lancet* 2007, 369(9569):1302-1313.

11. Statens Helsetilsyn: Psykiatriske poliklinikker. En evaluering av arbeidsformer og produktivitet innen voksen- og barne- og ungdomspsykiatrien. In.: Statens Helsetilsyn, Oslo; 2000.

12. Turi A, Bals M, Skre I, Kvernmo S: Health service use in indigenous Sami and non-indigenous youth in North Norway: A population based survey. *BMC Public Health* 2009, 9(1):378.

13. Helsedirektoratet: Veileder for poliklinikker i psykisk helsevern for barn og unge. In. Edited by Avdeling for psykisk helse. Oslo; 2008.

14. Heyerdahl S, Kvernmo S, Wichstrøm L: Self-reported behavioural/emotional problems in Norwegian adolescents from multiethnic areas. *European Child & Adolescent Psychiatry* 2004, 13(2):64-72.

15. American Psychological Association Presidential Task Force on Evidence-Based Practice: Evidence-based practice in psychology. *The American psychologist* 2006, 61(4):271-285.

16. Laird B, Smith B, Dutu G, Mellsop G: Views and experiences of family/whanau carers of psychiatric service users on diagnosis and classification. *The International journal of social psychiatry* 2010, 56(3):270-279.

17. Goin-Kochel RP, Mackintosh VH, Myers BJ: How many doctors does it take to make an autism spectrum diagnosis? *Autism* 2006, 10(5):439-451.

18. Siklos S, Kerns KA: Assessing the diagnostic experiences of a small sample of parents of children with autism spectrum disorders. *Research in Developmental Disabilities* 2007, 28(1):9-22.

19. Andersson HW, Ådnanes M, Hatling T: Nasjonal karlegging av tilbud om diagnostisering og helhetlig behandling av barn og ungdom med hyperkinetiske forstyrrelser/ADHD. Trondheim, Norway: SINTEF Helse; 2004.

20. Costello EJ, Egger H, Angold A: 10-Year Research Update Review: The Epidemiology of Child and Adolescent Psychiatric Disorders: I. Methods and Public Health Burden. *JAACAP* 2005, 44(10):972-986.

21. McClellan JM, Werry JS: Introduction--research psychiatric diagnostic interviews for children and adolescents. *J Am Acad Child Adolesc Psychiatry* 2000, 39(1):19-27.

22. Noda AM, Kraemer HC, Yesavage JA, Periyakoil VS: How many raters are needed for a reliable diagnosis? *International Journal of Methods in Psychiatric Research* 2001, 10(3):119-125.

23. Angold A, Costello EJ: Nosology and measurement in child and adolescent psychiatry. *J Child Psychol Psychiatry* 2009, 50(1-2):9-15.

24. Bjelland I, Dahl AA: Dimensjonal diagnostikk--ny klassifisering av psykiske lidelser. *Tidsskr Nor Laegeforen* 2008, 128(13):1541-1543.

25. Garb HN: Clinical judgment and decision making. *Annu Rev Clin Psychol* 2005, 1:67-89.

26. Rettew DC, Lynch AD, Achenbach TM, Dumenci L, Ivanova MY: Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research* 2009, 18(3):169-184.

27. Garland AF, Lau AS, Yeh M, McCabe KM, Hough RL, Landsverk JA: Racial and ethnic differences in utilization of mental health services among high-risk youths. *The American journal of psychiatry* 2005, 162(7):1336-1343.

28. Lewin A, Piacentini J: Evidence-Based Assessment of Child Obsessive Compulsive Disorder: Recommendations for Clinical Practice and Treatment Research. *Child and Youth Care Forum* 2010, 39(2):73-89.

29. Vaglum P: Er klinisk forskning farlig for klinisk praksis? *Tidsskr Nor Laegeforen* 2004, 124(15):1954-1955.

30. Bird HR: Global measures of impairment for epidemiologic and clinical use with children and adolescents. *International Journal of Methods in Psychiatric Research* 1996, 6(295-307):295-307.

31. Roberts RE: Prevalence of psychopathology among children and adolescents. *American Journal of Psychiatry* 1998, 155(6):715-725.

32. Ezpeleta L, de la Osa N, Domenech JM, Navarro JB, Losilla JM, Judez J: Diagnostic agreement between clinicians and the Diagnostic Interview for Children and Adolescents--DICA-R--in an outpatient sample. *J Child Psychol Psychiatry* 1997, 38(4):431-440.

33. Lauth B, Levy SR, Juliusdottir G, Ferrari P, Petursson H: Implementing the semi-structured interview Kiddie-SADS-PL into an in-patient adolescent clinical setting: impact on frequency of diagnoses. *Child Adolesc Psychiatry Ment Health* 2008, 2(1):14.

34. Spengler PM: The meta-analysis of clinical judgment project: Effects of experience on judgment accuracy. *The Counseling Psychologist* 2009, 37(350-399):350-399.

35. Ramirez Basco M, Bostic JQ, Davies D, Rush AJ, Witte B, Hendrickse W, Barnett V: Methods to improve diagnostic accuracy in a community mental health setting. *Am J Psychiatry* 2000, 157(10):1599-1605.

36. Miller PR, Dasher R, Collins R, Griffiths P, Brown F: Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews. *Psychiatry Research* 2001, 105(3):255-264.

37. Miller PR: Inpatient diagnostic assessments: 2. Interrater reliability and outcomes of structured vs. unstructured interviews. *Psychiatry Res* 2001, 105(3):265-271.

38. Williams JW, Jr., Noel PH, Cordes JA, Ramirez G, Pignone M: Is this patient clinically depressed? *JAMA* 2002, 287(9):1160-1170.

39. Foreman D, Morton S, Ford T: Exploring the clinical utility of the Development And Well-Being Assessment (DAWBA) in the detection of hyperkinetic disorders and associated diagnoses in clinical practice. *J Child Psychol Psychiatry* 2009, 50(4):460-470.

40. WHO: International Statistical Classification of Diseases and Related Health Problems, 10th revision. Geneva, Switzerland: Word Health Organization; 1992-94.

41. Reed GM, Ayuso-Mateos JL: Towards a more clinically useful International World Health Organisation classification of Mental Disorders. *Revista de Psiquiatria y Salud Mental* 2011(4):113-116.

42. Hyman SE: The diagnosis of mental disorders: the problem of reification. *Annu Rev Clin Psychol* 2010, 6:155-179.

43. Jewell J, Handwerk M, Almquist J, Lucas C: Comparing the validity of clinician-generated diagnosis of conduct disorder to the diagnostic interview schedule for children. *J Clin Child Adolesc Psychol* 2004, 33(3):536-546.

44. Tenney NH, Schotte CK, Denys DA, van Megen HJ, Westenberg HG: Assessment of DSM-IV personality disorders in obsessive-compulsive disorder: comparison of clinical diagnosis, self-report questionnaire, and semi-structured interview. *J Pers Disord* 2003, 17(6):550-561.

45. Dunerfeldt M, Elmlund A, Söderström B: Bedömningsinstrument inom BUP i Stockholm. Kartlägging och faktasammanställning. In. Edited by Barn- och ungdomspsykiatri, Stockholms läns landsting, Stockholm; 2010.

46. Andrews G, Anderson TM, Slade T, Sunderland M: Classification of anxiety and depressive disorders: problems and solutions. *Depression and anxiety* 2008, 25(4):274-281.

47. Pelham JWE, Fabiano GA, Massetti GM: Evidence-Based Assessment of Attention Deficit Hyperactivity Disorder in Children and Adolescents. *Journal of Clinical Child & Adolescent Psychology* 2005, 34(3):449-476.

48. Hudziak JJ, Achenbach TM, Althoff RR, Pine DS: A dimensional approach to developmental psychopathology. *Int J Methods Psychiatr Res* 2007, 16 Suppl 1:S16-23.

49. Kraemer HC: DSM categories and dimensions in clinical and research contexts. *International Journal of Methods in Psychiatric Research* 2007, 16(S1):S8-S15.

50. Verhulst FC, van der Ende J: Agreement between parents' reports and adolescents' self-reports of problem behavior. *J Child Psychol Psychiatry* 1992, 33(6):1011-1023.

51. Kraemer HC, Measelle JR, Ablow JC, Essex MJ, Boyce WT, Kupfer DJ: A new approach to integrating data from multiple informants in psychiatric assessment and research: mixing and matching contexts and perspectives. *Am J Psychiatry* 2003, 160(9):1566-1577.

52. Noordhof A, Oldehinkel AJ, Verhulst FC, Ormel J: Optimal use of multi-informant data on co-occurrence of internalizing and externalizing problems: the TRAILS study. *International Journal of Methods in Psychiatric Research* 2008, 17(3):174-183.

53. Rutter M: Research review: Child psychiatric diagnosis and classification: concepts, findings, challenges and potential. *Journal of child psychology and psychiatry, and allied disciplines* 2011, 52(6):647-660.

54. Andrews G, Goldberg DP, Krueger RF, Carpenter WT, Hyman SE, Sachdev P, Pine DS: Exploring the feasibility of a meta-structure for DSM-V and ICD-11: could it improve utility and validity? *Psychol Med* 2009, 39(12):1993-2000.

55. Kendler KS, Prescott CA, Myers J, Neale MC: The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. *Arch Gen Psychiatry* 2003, 60(9):929-937.

56. O'Connor BP: The search for dimensional structure differences between normality and abnormality: a statistical review of published data on personality and psychopathology. *Journal of personality and social psychology* 2002, 83(4):962-982.

57. Krueger RF, Chentsova-Dutton YE, Markon KE, Goldberg D, Ormel J: A cross-cultural study of the structure of comorbidity among common psychopathological syndromes in the general health care setting. *J Abnorm Psychol* 2003, 112(3):437-447.

58. Belfer ML: Child and adolescent mental disorders: the magnitude of the problem across the globe. *J Child Psychol Psychiatry* 2008, 49(3):226-236.

59. Staller JA: Diagnostic profiles in outpatient child psychiatry. *Am J Orthopsychiatry* 2006, 76(1):98-102.

60. Mathiesen KS, Karevold E, Knudsen AK: Psykiske lidelser blant barn og unge i Norge. Oslo: Nasjonalt folkehelseinstitutt; 2009.

61. Heiervang E, Stormark KM, Lundervold AJ, Heimann M, Goodman R, Posserud MB, Ullebo AK, Plessen KJ, Bjelland I, Lie SA *et al*: Psychiatric disorders in Norwegian 8- to 10-year-olds: an epidemiological survey of prevalence, risk factors, and service use. *JAACAP* 2007, 46(4):438-447.

62. Helse og Omsorgsdepartementet: St.prp.nr. 63. Om opptrappingsplan for psykisk helse 1999 - 2006. Endringer i statsbudsjettet for 1998. Oslo; 1998.

63. WHO: Towards a common language for functioning, disability and health: ICF. Geneva: World Health Organisation; 2002.

64. Mathai J, Anderson P, Bourne A: Comparing psychiatric diagnoses generated by the Strengths and Difficulties Questionnaire with diagnoses made by clinicians. *Aust N Z J Psychiatry* 2004, 38(8):639-643.

65. York A, Lamb C: Building and Sustaining Specialist CAMHS. Workforce, Capacity and Functions of Tiers 2, 3 and 4 Specialist Child and Adolescent Mental Health Services Across England, Ireland, Northern Ireland, Scotland and Wales. In.: Royal College of Psychiatrists; 2005.

66. Kohn R, Saxena S, Levav I, Saraceno B: The treatment gap in mental health care. *Bull World Health Organ* 2004, 82(11):858-866.

67. Kowatch RA, Youngstrom EA, Danielyan A, Findling RL: Review and meta-analysis of the phenomenology and clinical characteristics of mania in children and adolescents. *Bipolar disorders* 2005, 7(6):483-496.

68. Achenbach TM, Rescorla LA: Multicultural Understanding of Child and Adolescent Psychopathology: Implications for Mental Health Assessment. *American Journal of Psychiatry* 2007, 164(6):983-984.

69. Akobeng AK: Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatr* 2007, 96(3):338-341.

70. Altman DG, Bland JM: Diagnostic tests. 1: Sensitivity and specificity. *Bmj* 1994, 308(6943):1552.

71. Glaros AG, Kline RB: Understanding the accuracy of tests with cutting scores: the sensitivity, specificity, and predictive value model. *J Clin Psychol* 1988, 44(6):1013-1023.

72. Akobeng AK: Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice. *Acta Paediatr* 2007, 96(4):487-491.

73. Warner J: Clinicians' guide to evaluating diagnostic and screening tests in psychiatry. *Adv Psychiatr Treat* 2004, 10(6):446-454.

74. Deeks JJ, Altman DG: Diagnostic tests 4: likelihood ratios. *Bmj* 2004, 329(7458):168-169.

75. Bhopal R: Concepts of epidemiology. Integrating the ideas, theories, principles and methods of epidemiology. Second edition. Oxford, New York: Oxford University Press; 2008.

76. Myers KM, Sulzbacher S, Melzer SM: Telepsychiatry with children and adolescents: are patients comparable to those evaluated in usual outpatient care? *Telemedicine journal and e-health : the official journal of the American Telemedicine Association* 2004, 10(3):278-285.

77. Urness D, Wass M, Gordon A, Tian E, Bulger T: Client acceptability and quality of life--telepsychiatry compared to in-person consultation. *J Telemed Telecare* 2006, 12(5):251-254.

78. Greenberg N, Boydell KM, Volpe T: Pediatric telepsychiatry in ontario: Caregiver and service provider perspectives. *The journal of behavioral health services & research* 2006, 33(1):105-111.

79. Martin-Khan M, Wootton R, Whited J, Gray LC: A systematic review of studies concerning observer agreement during medical specialist diagnosis using videoconferencing. *J Telemed Telecare* 2011, 17(7):350-357.

80. Pesamaa L, Ebeling H, Kuusimaki ML, Winblad I, Isohanni M, Moilanen I: Videoconferencing in child and adolescent telepsychiatry: a systematic review of the literature. *J Telemed Telecare* 2004, 10(4):187-192.

81. Hanssen B, Wangberg SC, Gammon D: Use of videoconferencing in Norwegian psychiatry. *J Telemed Telecare* 2007, 13(3):130-135.

82. Diamond JM, Bloch RM: Telepsychiatry assessments of child or adolescent behavior disorders: a review of evidence and issues. *Telemedicine journal and e-health : the official journal of the American Telemedicine Association* 2010, 16(6):712-716.

83. Kuhn C, Winkler Metzke C, Aebi M, Steinhausen HC: PW01-59 - Effects of an internet based assessment of child and adolescent psychopathology (DAWBA) on clinical decision making. *European Psychiatry* 2010, 25, Supplement 1(0):1475.

84. Amstadter AB, Broman-Fulks J, Zinzow H, Ruggiero KJ, Cercone J: Internet-based interventions for traumatic stress-related mental health problems: A review and suggestion for future research. *Clinical Psychology Review* 2009, 29(5):410-420.

85. Spek V, Cuijpers P, Nyklicek I, Riper H, Keyzer J, Pop V: Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. *Psychol Med* 2007, 37(03):319-328.

86. Ford T, Goodman R, Meltzer H: The British Child and Adolescent Mental Health Survey 1999: The Prevalence of DSM-IV Disorders. *JAACAP* 2003, 42(10):1203-1211.

87. Meltzer H, Gatward R, Goodman R, Ford T: Mental health of children and adolescents in Great Britain. *International Review of Psychiatry* 2003, 15(1-2):185-187.

88. Frigerio A, Vanzin L, Pastore V, Nobile M, Giorda R, Marino C, Molteni M, Rucci P, Ammaniti M, Lucarelli L *et al*: The Italian preadolescent mental health project (PrISMA): rationale and methods. *Int J Methods Psychiatr Res* 2006, 15(1):22-35.

89. Jensen-Doss A, Weisz JR: Diagnostic agreement predicts treatment process and outcomes in youth mental health clinics. *J Consult Clin Psychol* 2008, 76(5):711-722.

90. Angold A: Diagnostic interviews with parents and children. In: *Child and Adolescent Psychiatry: Modern Approaches, 4th ed.* edn. Edited by Rutter M, Taylor E. Oxford: Blackwell Scientific; 2002: 32-51.

91. Haine R, Brookman-Frazee L, Tsai K, Roesch S, Garland A: Clinician Perspectives of Diagnosis and Perceived Client Change in "Real World" Psychotherapy for Youth Emotional and Behavioral Disorders. *Journal of Child and Family Studies* 2007, 16(5):712-728.

92. DAWBA information for researchers and clinicians about the Development and Weel-Being Assessment [www.dawba.info]

93. Hanssen-Bauer K, Aalen OO, Ruud T, Heyerdahl S: Inter-rater reliability of clinician-rated outcome measures in child and adolescent mental health services. *Adm Policy Ment Health* 2007, 34(6):504-512.

94. Hanssen-Bauer K, Gowers S, Aalen OO, Bilenberg N, Brann P, Garralda E, Merry S, Heyerdahl S: Cross-national reliability of clinician-rated outcome measures in child and adolescent mental health services. *Adm Policy Ment Health* 2007, 34(6):513-518.

95. Organization WH: The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research. Geneva: WHO; 1993.

96. Organization WH: Classification of Mental and Behavioural Disorders, Clinical Description and Diagnostic Guidelines. Geneva: WHO; 1992.

97. Goodman A, Heiervang E, Collishaw S, Goodman R: The 'DAWBA bands' as an ordered-categorical measure of child mental health: description and validation in British and

Norwegian samples. *Social psychiatry and psychiatric epidemiology* 2011, 46(6):521-532.

98. Goodman R, Ford T, Richards H, Gatward R, Meltzer H: The Development and Well-Being Assessment: Description and Initial Validation of an Integrated Assessment of Child and Adolescent Psychopathology. *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 2000, 41(5):645-655.

99. Heiervang E, Goodman A, Goodman R: The Nordic advantage in child mental health: separating health differences from reporting style in a cross-cultural comparison of psychopathology. *J Child Psychol Psychiatry* 2008, 49(6):678-685.

100. Foreman DM, Ford T: Assessing the diagnostic accuracy of the identification of hyperkinetic disorders following the introduction of government guidelines in England. *Child Adolesc Psychiatry Ment Health* 2008, 2(1):32.

101. Goodman R: The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry, and allied disciplines* 1997, 38(5):581-586.

102. Goodman R: The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. *Journal of child psychology and psychiatry, and allied disciplines* 1999, 40(5):791-799.

103. Goodman R, Renfrew D, Mullick M: Predicting type of psychiatric disorder from Strengths and Difficulties Questionnaire (SDQ) scores in child mental health clinics in London and Dhaka. *European child & adolescent psychiatry* 2000, 9(2):129-134.

104. Pirkis J, Burgess P, Coombs T, Clarke A, Jones-Ellis D, Dickson R: Routine measurement of outcomes in Australia's public sector mental health services. *Aust New Zealand Health Policy* 2005, 2(1):8.

105. Schorre BE, Vandvik IH: Global assessment of psychosocial functioning in child and adolescent psychiatry. A review of three unidimensional scales (CGAS, GAF, GAPD). *Eur Child Adolesc Psychiatry* 2004, 13(5):273-286.

106. Green B, Shirk S, Hanze D, Wanstrath J: The Children's Global Assessment Scale in Clinical Practice: An Empirical Evaluation. *JAACAP* 1994, 33(8):1158-1164.

107. Dyrborg J, Larsen FW, Nielsen S, Byman J, Nielsen BB, Gautre-Delay F: The Children's Global Assessment Scale (CGAS) and Global Assessment of Psychosocial Disability (GAPD) in clinical practice--substance and reliability as judged by intraclass correlations. *Eur Child Adolesc Psychiatry* 2000, 9(3):195-201.

108. Bird HR, Canino G, Rubio-Stipec M, Ribera JC: Further measures of the psychometric properties of the Children's Global Assessment Scale. *Arch Gen Psychiatry* 1987, 44(9):821-824.

109. Gowers SG, Harrington RC, Whitton A, Lelliott P, Beevor A, Wing J, Jezzard R: Brief scale for measuring the outcomes of emotional and behavioural disorders in children. Health of the Nation Outcome Scales for children and Adolescents (HoNOSCA). *Br J Psychiatry* 1999, 174:413-416.

110. Pirkis JE, Burgess PM, Kirk PK, Dodson S, Coombs TJ, Williamson MK: A review of the psychometric properties of the Health of the Nation Outcome Scales (HoNOS) family of measures. *Health Qual Life Outcomes* 2005, 3:76.

111. Garralda ME, Yates P, Higginson I: Child and adolescent mental health service use: HoNOSCA as an outcome measure. *Br J Psychiatry* 2000, 177(1):52-58.

112. Brann P, Coleman G, Luk E: Routine outcome measurement in a child and adolescent mental health service: an evaluation of HoNOSCA. *Australian and New Zealand Journal of Psychiatry* 2001, 35(3):370-376.

113. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM: Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008, 149(12):889-897.

114. Cicchetti DV, Sparrow SA: Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic* 1981, 86(2):127-137.

115. McGraw KO: Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1996, Electronic(30-46):30-46.

116. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979, 86(2):420-428.

117. Shrout PE: Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research* 1999, 7(3):301-317.

118. Fischer JE, Bachmann LM, Jaeschke R: A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med* 2003, 29(7):1043-1051.

119. Helmen Borge AI: Psykologi og forskningsetikk: Kan deltakelse i forskningsprosjekt gi psykiske skader? In: *Forskningsetikk Beskyttelse av enkeltpersoner og samfunn.* Edited by Ruyter KW. Oslo, Norway: Gyldendal; 2003.

120. Arnold LE, Stoff DM, Cook Jr E, Cohen DJ, Kruesi M, Wright C, Hattab J, Graham P, Zametkin A, Castellanos FX *et al*: Ethical Issues in Biological Psychiatric Research with Children and Adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry* 1995, 34(7):929-939.

121. Ot.prp. nr. 74: Om lov om medisinsk og helsefaglig forskning (Helseforskningsloven). In.; 2006-2007.

122. The World Medical Association: WMA Declaration of Helsinki - Ethical principles for medical research involving human subjects. In.; 2008.

123. Vitiello B, Jensen PS, Hoagwood K: Integrating science and ethics in child and adolescent psychiatry research. *Biol Psychiatry* 1999, 46(8):1044-1049.

124. Forsberg E-M: Verktøykasse for etiske vurderinger. In: *Forskningsetikk Beskyttelse av enkeltpersoner og samfunn.* Edited by Ruyter KW. Oslo, Norway: Gyldendal; 2003.

125. Morrow V, Richards M: The Ethics of Social Research with Children: An Overview1. *Children & Society* 1996, 10(2):90-105.

126. Ruyter KW: Det informerte samtykket i medisinsk forskning - mellom selvråderett og beskyttelse. In: *Forskningsetikk Beskyttelse av enkeltpersoner og samfunn.* Edited by Ruyter KW. Oslo, Norway: Gyldendal; 2003.

127. Smith-Tyler J: Informed consent, confidentiality, and subject rights in clinical trials. *Proceedings of the American Thoracic Society* 2007, 4(2):189-193; discussion 193.

128. Mason SA, Allmark PJ: Obtaining informed consent to neonatal randomised controlled trials: interviews with parents and clinicians in the Euricon study. *Lancet* 2000, 356(9247):2045-2051.

129. Tobias JS, Souhami RL: Fully informed consent can be needlessly cruel. *Bmj* 1993, 307(6913):1199-1201.

130. Backe-Hansen E: Barn. In., vol. 2012. http://etikkom.no/no/FBIB/Temaer/Forskning-pa-bestemte-grupper/Barn/; 2009.

131. Kumra S, Briguglio C, Lenane M, Goldhar L, Bedwell J, Venuchekov J, Jacobsen LK, Rapoport JL: Including children and adolescents with schizophrenia in medication-free research. *Am J Psychiatry* 1999, 156(7):1065-1068.

132. Gilbody SM, Song F: Publication bias and the integrity of psychiatry research. *Psychol Med* 2000, 30(2):253-258.

133. Gilbody SM, Song F, Eastwood AJ, Sutton A: The causes, consequences and detection of publication bias in psychiatry. *Acta Psychiatr Scand* 2000, 102(4):241-249.

134. Hellevik O: Forskningsmetode i sosiologi og statsvitenskap. Oslo, Norway: Universitetsforlaget; 1989.

135. Sogaard AJ, Selmer R, Bjertness E, Thelle D: The Oslo Health Study: The impact of self-selection in a large, population-based survey. *International journal for equity in health* 2004, 3(1):3.

136. Vink JM, Willemsen G, Stubbe JH, Middeldorp CM, Ligthart RS, Baas KD, Dirkzwager HJ, de Geus EJ, Boomsma DI: Estimating non-response bias in family studies: application to mental health and lifestyle. *European journal of epidemiology* 2004, 19(7):623-630.

137. Pedersen PB, Sitter M, Lilleeng SE: Pasienter i det psykiske helsevernet 2009. Edited by Helsedirektoratet. Oslo, Norway; 2011.

138. Lundh A, Kowalski J, Sundberg CJ, Gumpert C, Landen M: Children's Global Assessment Scale (CGAS) in a naturalistic clinical setting: Inter-rater reliability and comparison with expert ratings. *Psychiatry Res* 2010, 177(1-2):206-210.

139. Angold A, Weissman MM, John K, Merikangas KR, Prusoff BA, Wickramaratne P, Gammon GD, Warner V: Parent and child reports of depressive symptoms in children at low and high risk of depression. *Journal of child psychology and psychiatry, and allied disciplines* 1987, 28(6):901-915.

140. Berg-Nielsen TS, Vika A, Dahl AA: When adolescents disagree with their mothers: CBCL-YSR discrepancies related to maternal depression and adolescent self-esteem. *Child Care Health Dev* 2003, 29(3):207-213.

141. Goodman R, Ford T, Simmons H, Gatward R, Meltzer H: Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *Br J Psychiatry* 2000, 177:534-539.

142. Hysing M, Elgen I, Gillberg C, Lie SA, Lundervold AJ: Chronic physical illness and mental health in children. Results from a large-scale population study. *Journal of child psychology and psychiatry, and allied disciplines* 2007, 48(8):785-792.

143. Angold A, Costello EJ, Farmer EM, Burns BJ, Erkanli A: Impaired but undiagnosed. *Journal of the American Academy of Child and Adolescent Psychiatry* 1999, 38(2):129-137.

144. Baca-Garcia E, Perez-Rodriguez MM, Basurte-Villamor I, Fernandez del Moral AL, Jimenez-Arriero MA, Gonzalez de Rivera JL, Saiz-Ruiz J, Oquendo MA: Diagnostic stability of psychiatric disorders in clinical practice. *The British journal of psychiatry : the journal of mental science* 2007, 190:210-216.

145. Kraemer HC, Kupfer DJ, Clarke DE, Narrow WE, Regier DA: DSM-5: How Reliable Is Reliable Enough? *The American journal of psychiatry* 2012, 169(1):13-15.

146. Hyman SE: Can neuroscience be integrated into the DSM-V? *Nat Rev Neurosci* 2007, 8(9):725-732.

147. Pies R: How "objective" are psychiatric diagnoses?: (guess again). *Psychiatry (Edgmont)* 2007, 4(10):18-22.

Paper 1

Paper 2

Paper 3

ISBN xxx-xx-xxxx-xxx-x