



Diagnosing Bipolar Disorder with the Diagnostic Interview for Psychoses (DIP).

A Reliability Study.

Caroline Spant Skorven

and

Guro Mikalsen

Supervisor: Ingunn Skre

Main Thesis for the Cand. Psychol. Degree

Department of Psychology

University of Tromsø

Spring 2010

Preface and Acknowledgements

This study is part of a larger research project that aims to assess the validity and reliability of a Norwegian translation of the Diagnostic Interview for Psychoses (DIP). We were introduced to the study by the project leader, who is also our supervisor, Ingunn Skre. The lack of diagnostic interviews directed specifically towards the most severe mental disorders inspired us to take part in the study, and the focus on bipolar disorder was chosen because it seems to be one of the most challenging disorders to diagnose. In addition to high rates of misdiagnosis, individuals suffering from this disorder are also at risk for delayed diagnosis, and faced with potentially devastating consequences of delayed or incorrect treatment. For these reasons, we have seen the need for better diagnostic tools that can accurately diagnose bipolar disorder and contribute to better care and support for the affected individuals.

Working with this paper has been both fun and challenging, but most of all a process of acquiring valuable knowledge about the diagnostic process, and about bipolar disorder and its impact on the lives of those affected by it. Caroline joined the research project in January 2009, working with translating the interview and with the literature base, and Guro has been involved since August 2009. Since then, we have both participated in the data collection, and worked together on acquiring literature, conducting the data analysis, and writing this paper.

Many have contributed to our work on this paper. First and foremost, we would like to thank those who participated as respondents in the study, who has not only given us the opportunity to assess the DIP, but also inspired us greatly by their interest in this field of research and in what we do. We are also grateful to the management and staff at the University Hospital for Northern Norway in Tromsø and Nordland Hospital in Bodø, for allowing us to conduct the study, and for their efforts in recruiting respondents and making the data collection process easier. This paper has also been made possible by the funding that the larger research project has received. Caroline was provided with a student research fellowship by the Department for Psychology at the University of Tromsø, that was granted by The Norwegian Research Council (NFR). This enabled her to begin her work with the translation process and on the literature base in January 2009. The project also received funding from the Northern Norway Regional Health Authority Research Council (Helse Nord RHF), for the data collection in the study of DIP.

We want to thank our supervisor, Ingunn Skre, for introducing us to the project, for her contributions by providing advice about relevant literature, the data analysis, and the structuring of this paper, and for sharing with us her expertise in clinical psychology and diagnostic processes. In terms of finding relevant literature, we also thank Terje Øiesvold for allowing us to use his collection of literature on bipolar disorder. We also thank Skre and Øiesvold, as well as Vidje Hansen, Lena-Kristin Nerdal and Oxana Nikiforova, for participating in the data collection by conducting interviews, and Connie Villemo Nilsen for her assistance with managing and organising the database.

The work on the Italian translation of the DIP served as guidance in this work on the Norwegian translation and thus, we are grateful to Alberto Rossi and Francesco Ammadeo for their efforts. We also thank professor Assen Jablensky and Vera Morgan for providing us with the original DIP and allowing a Norwegian translation to be made. We also acknowledge the work of Peter McGuffin and Ann Farmer as the originators of the OPCRIT checklist and diagnostic algorithm, which is the foundation of the DIP.

Abstract

A reliability study of the Norwegian translation of the Diagnostic Interview for Psychoses (DIP) was conducted, with a focus on its usefulness when diagnosing bipolar disorder. Emphasis is given to the severity and importance of correct diagnosis and treatment of this disorder. The DIP is a semi-structured interview that aims to assess low prevalent psychiatric disorders. It is based on the Operational Criteria for Psychoses (OPCRIT), and generates diagnoses according to several diagnostic systems. The respondents included in the study were inpatients at the psychiatric hospitals in Tromsø and Bodø. Twenty-seven respondents were independently assessed by two raters, one of whom conducted the interview, while the other scored from observation. Fourteen of them were re-interviewed later by a third independent rater. *Inter-rater reliability* was generally high, with good to excellent agreement on the majority of items. Agreement on broad diagnostic categories was also high. *Test-retest reliability* for both individual items and broad diagnostic categories ranged from moderate to excellent. The findings are consistent with the results of previous studies of the reliability of DIP. The results reported here are based on preliminary data from a larger research project, and should be viewed with some caution. The findings and their possible implications are discussed in relation to bipolar disorder and the conceptualisation of this, and other, severe mental disorders.

In recent years there has been increased attention to the importance of correct diagnosis and treatment of severe mental illness. However, there is no complete international consensus on the definitions of severe mental disorders (Lora, Bezzi, & Erlicher, 2007; Ruggeri, Leese, Thornicroft, Bisoffi, & Tansella, 2000). Common criteria, however, seems related to both diagnosis and severity, where the latter is assessed in terms of recent treatment, symptoms, and functioning in social and occupational domains. The low consistency of definitions lead to difficulty in estimating the prevalence of severe mental illnesses as a whole, and of the individual disorders. It also reflects the fact that identifying severe mental disorders, and distinguishing them from each other, can be very challenging both in clinical and research settings, especially when it comes to low prevalent disorders. As we will see, bipolar disorder is a disorder that has proven hard to diagnose, much due to its shared symptoms with other low prevalent disorders, such as schizophrenia, schizoaffective disorder and major depressive disorder, and due to the difficulties associated with detecting its defining symptoms. Nevertheless, reaching agreement between professionals may be easier on the level of individual symptoms, than on complete diagnostic categories or syndromes. While the definition of what constitutes a disorder may change as research progress, the definitions of single symptoms and illness signs seems to be more stable.

Both in clinical practice and research, diagnostic interviews are commonly used to assess, and distinguish between, different psychiatric disorders. However, there exist few comprehensive diagnostic tools directed specifically towards low prevalent disorders. Furthermore, most existing interviews are constructed to make diagnostic decisions according to one of the major diagnostic systems, currently the International Classification of Diseases, ICD-10 (World Health Organization, 1993), and the Diagnostic and Statistical manual of Mental Disorders, DSM-IV (American Psychiatric Association, 1994). When these systems are revised, which they are every decade or so, diagnoses made according to these interviews are not always easy to convert to the new system. For these reasons, the Diagnostic Interview for Psychoses (DIP), was developed specifically for the Australian National Mental Health Survey - Low Prevalence (Psychotic) Disorders Study, conducted in 1997 and 1998 (Castle et al., 2006; Jablensky et al., 2000). The diagnostic module of the DIP (DIP-DM, see appendix A for an overview) is based on the Operational Criteria For Psychosis (OPCRIT) 90 item check list (McGuffin, Farmer, & Harvey, 1991; Williams et al., 1996). The merit of the OPCRIT is that it registers all well known symptoms and signs of severe mental disorder, and

this registration is not influenced by the current definitions of the disorders in specific diagnostic systems. Thus, the use of OPCRIT and interviews such as the DIP, opens up for a dimensional approach to mental disorders. The information obtained in the DIP can subsequently be entered into the OPCRIT algorithm, and yield diagnoses according to any diagnostic system, including the ICD-10 (World Health Organization, 1993) and the three latest versions of the DSM (American Psychiatric Association, 1980; 1987; 1994) (see appendix B for a list of all diagnostic classification systems compatible with DIP). DIP-DM is accompanied by computer software that serves this purpose. While the DIP-DM can be used alone when appropriate, the complete DIP also incorporates important areas other than the purely symptomatic, such as social functioning and disability, and patterns of service utilization.

Assessment of the reliability and validity of the English DIP has shown good results in Australia, where the interview origins, and the DIP has been considered useful both in terms of more accurate diagnosis of psychotic disorders and for use in epidemiological research (Castle et al., 2006). This was supported by a study of an Italian translation of the DIP (Rossi et al., 2010). Accordingly, the DIP may prove useful in making more accurate diagnoses of bipolar disorder, and other low prevalent mental disorders, in Norwegian psychiatric health care as well. For this reason, our study aims to test the reliability of a Norwegian translation of the DIP-DM, in a Norwegian patient population. We have chosen a focus on bipolar disorder, as this is a low prevalent disorder that has proved hard to diagnose, and that seems to be one of the severe mental disorders most vulnerable to misdiagnosis or delayed diagnosis (e.g. Øiesvold et al., submitted). Thus, in addition to describing the reliability study of the DIP, we emphasize the extent and consequences of inadequate diagnosing of bipolar disorder.

Bipolar disorder

Epidemiological studies have indicated a fairly low prevalence rate of bipolar disorder, ranging from 1 to 2% (Angst, 2007; Glick, 2004; Hirschfeld, 2002; Kessler, et al. 1994; Piver, Yatham, & Lam, 2002; Regier et al., 1988). Estimates of prevalence reported from the Norwegian population are comparable to this, as the Norwegian department of health care report a prevalence of 1% (NOU (Norwegian Governmental Report), 1999), and a

study of DSM-III-R axis I disorders in Oslo showed a lifetime prevalence of bipolar disorder of 1.6% (Torgersen, Cramer, & Kringlen, 2002).

Though bipolar disorder has low prevalence, the impact of the disorder on the lives of the affected individuals and their families is considerable, and may also represent a substantial strain on health care resources. In addition to the general impact severe mental illness has on quality of life, the mortality rate among bipolar patients is high. An indirect contributor to the high rate is psychotic or reckless behaviour leading to accidents and drug abuse. However, the main cause of the high mortality rate among bipolar patients is suicide, and research has shown that the lifetime risk of suicide among these patients is 10 to 20 times higher than in the general population (e.g. Mork, Mehlum, & Walby, 2009; Tondo, Isacsson, & Baldessarini, 2003; Ösby, Brandt, Correia, Ekborn, & Sparén, 2001).

Without treatment, manic episodes have an average duration of 4 to 6 months, and depressive episodes last for 6 to 9 months, on average. With treatment, the duration of acute illness periods can be dramatically decreased, and the total illness time may be reduced with as much as 50 to 75% (Sachs & Thase, 2000). There is strong evidence of genetic and biological factors related to bipolar disorder (e.g. Edvardsen et al., 2008), and treatment with psychopharmaceuticals is almost always necessary. Because the treatment is aimed both at reducing symptoms in acute phases, and subsequently at preventing relapses, individuals treated for bipolar disorder will most likely need relatively frequent contact with health care services. In addition, psychosocial treatment, such as psychoeducation and psychotherapy, is very important in preventing relapses and exacerbation of episodes, as it may enable patients to better manage their illness and to seek medical care when appropriate. Thus, the severe negative consequences that bipolar disorder may have, and the importance of correct treatment to reduce these, makes it highly important that individuals suffering from this disorder are diagnosed and given adequate treatment as soon as possible.

Diagnosing bipolar disorder

Today, the most common diagnostic systems are the ICD-10 and the DSM-IV, the first being the most commonly used worldwide, while the latter is the most used in North-America. These two systems operate with somewhat diverging definitions of bipolar disorder. The DSM-IV system requires at least one full-blown episode of mania for the Bipolar I diagnosis to be given, and at least both a major depressive and a hypomanic episode for the

Bipolar II diagnosis to be given. ICD-10 defines bipolar disorder by at least two episodes of altered mood state and activity level. Consequently, the ICD-10 system does not differentiate between bipolar I and bipolar II, and as long as there has been at least two hypomanic episodes, no major depressive episode is required for a diagnosis of bipolar disorder. Thus, the ICD-10 system can be said to define bipolar disorder in a broader manner than the DSM-IV (Farmer, Wessley, Castle, & McGuffin, 1992), and patients with the same symptomatic picture can fall within different diagnostic groups, depending on the system being used.

The diagnostic systems differentiates between several different types of episodes (see appendix C for an overview of sub-classification of bipolar disorder), but in general, both systems have been criticized for their definition of bipolar disorder, in terms of their usefulness in objectively distinguishing between diagnostic categories and level of severity, and detection of subclinical or atypical cases (Craddock, Jones, Kirov, & Jones, 2004). Akiskal (2008) argue that diagnostic systems should operate with a bipolar spectrum, and include sub-threshold conditions to help identify individuals with less severe forms of bipolar disorder, that might be in need of treatment. As a result, an expanded concept of bipolar spectrum disorders has been defined (e.g. Marneros & Angst, 2000), though as the features of these subtypes are not currently defined in the diagnostic systems, they may only be considered as bipolar disorder not otherwise specified (Piver, Yatham, & Lam, 2002).

In addition to criticism of definitions of individual disorders, it has been suggested that the categorical approach of current diagnostic systems may be insufficient, and that it should at least be supplemented by a more dimensional approach (e.g. Widiger & Samuel, 2005). More specifically, the validity of the current distinction between major affective disorders, schizoaffective disorders, and schizophrenia, has been questioned, and the criteria used to define these disorders today considered insufficient (e.g. Laursen, Agerbo, & Pedersen, 2009; Marneros, 2003). Adding to the ongoing work on how to conceptualise these severe mental disorders, research has shown that while there are differences between bipolar disorder and schizophrenia, there are also similarities in terms of epidemiological (e.g. Torrey, 1999) and genetic (e.g. Lichtenstein et al., 2009; Van Snellenberg & Candia, 2009) factors. This raises questions about the validity of the current definition of schizoaffective disorder in the diagnostic systems. It may be that disorders such as bipolar disorder, schizoaffective disorder, and schizophrenia, should be viewed more as part of a continuum, rather than as distinct entities. As Buckley and colleagues (2004) concluded in their review of

the subject, there is no conclusive evidence about the relationship between these disorders, but it seems unlikely that they are completely unrelated. Thus, considering other factors than those currently specified in the diagnostic systems, may prove useful in detecting these disorders, and in providing appropriate treatment and support for patients affected by them.

The difficulties in detecting bipolar disorder is reflected by the fact that while the disorder normally has its debut early in life, around the age of 20 (Kringlen, 2002; Schulze, Hedeker, Zandi, Rietschel, & McMahon, 2006), getting the correct diagnosis can be a lengthy process. Studies have shown that for a significant number of patients, the time elapsed between occurrence of the first symptoms and correct diagnosis can be several years (Hirschfeld, Lewis, & Vornik, 2003; Lish, Dime-Meenan, Whybrow, Price, & Hirschfeld, 1994; Suppes et al., 2001). Lish and colleagues (1994) found that 48% of patients saw three or more professionals before being diagnosed with bipolar disorder, while 10% saw seven or more. For 34% of their respondents, more than a decade passed between first contact with health service professionals and a diagnosis of bipolar disorder. In addition, Hirschfeld and colleagues (2003) reported that 69% of their respondents claimed to have been misdiagnosed at some point, and the average was to receive 3.5 other diagnoses before one of bipolar disorder.

Challenges related to diagnosing bipolar disorder lies with both patients and health service professionals. On the part of the patient there are issues concerning the information they provide to health service professionals. It can be difficult for them to recognize that what they are experiencing are symptoms of mental disorder, both because of the nature of the disorders (making them feel good about themselves, rather than ill, in manic episodes) and the poorer knowledge in the population of these disorders relative to more common disorders such as depression or pathological anxiety. The lack of insight by patients in acute manic episodes, even after treatment and improvement of other symptoms, was demonstrated by Ghaemi, Stoll, and Pope (1995), with the use of a semi-structured interview assessing patients recognition of symptoms of illness, and the extent of patients recognition of need for treatment with hospitalisation, medications, and psychiatric follow-up. Patients might also be reluctant to disclose symptoms of low prevalent disorders to others due to fear of stigma, or due to delusional beliefs about the consequences of reporting their experiences.

Hirschfeld (2001) described the treatment-seeking behaviour of patients with bipolar disorder, in a report using data collected from the members of the National Depressive and

Manic-Depressive Association (DMDA), and showed that 60% of patients sought treatment only in depressive states. The percentage might be even higher in populations including patients with unrecognised bipolar disorder, as members of the DMDA may be expected to have greater knowledge of their disorder and need for treatment. This pattern of treatment-seeking behaviour, and patients' lack of insight, also contributes to make the cooperation with health service providers difficult, because of the potential discrepancy between professionals and patients view of their state. Consequently, at least some patients will not make contact with health services until someone in their social network takes action on the basis of changes in their behaviour, or until they endanger themselves or others in ways that warrant hospital admission. Thus, while patients with bipolar disorder may seek help voluntarily in depressive states, their contact with the health care system in manic phases often involves involuntary care. Patients in hypomanic phases very rarely seek help, because they feel good about themselves, and their behaviour is not deviant enough to justify commitment to psychiatric hospital.

The challenges to accurate diagnosis of bipolar disorder are consequently related to the knowledge and diagnostic methods used amongst health service professionals. When meeting the bipolar patient in a depressive state, correct diagnosis depends on the professional's skills and use of structured diagnostic approaches in asking for past manic or hypomanic episodes. Studies have indicated that clinicians failure to detect symptoms of mania may be due to a tendency not to ask beyond the currently presenting symptoms, especially when these are depressive (Taiminen et al., 2001; Øiesvold et al., submitted). Because of this, and the treatment-seeking behaviour of patients with bipolar disorder, periodical variations in mood may never be discussed. Furthermore, if symptoms of mania are identified, the manifestations of the disorder can be confusing, especially in patients with depressive and manic mixed states. Swann, Steinberg, Lijffijt, and Moeller (2009) demonstrated that patients in mixed states are more likely to simultaneously experience psychosis and anxiety than patients in pure depressive or manic states, making the clinical picture more complex. Thus, symptoms may not always be attributed to bipolar disorder, because symptoms overlap with several other severe mental disorders and can be mistaken for these. At the same time, comorbidity is common in patients with bipolar disorder, and this complicates both the diagnosing and course of the disorder (Goodwin & Jamison, 2007). The use of structured diagnostic interviews, which explicitly demands information about different

types of illness episodes, and about other factors known to be related to specific disorders, may improve the detection of manic symptoms and contribute to distinguish severe mental disorders from each other.

Research has identified alternate diagnoses often received by patients with bipolar disorder (e.g. Hirschfeld et al., 2003; Lish et al., 1994). Misdiagnosis as unipolar depression is by far the most commonly reported in such studies, and diagnoses of schizophrenia, schizophreniform disorder, schizoaffective disorder, emotionally unstable or antisocial personality disorder, anxiety disorders, and alcohol or substance abuse is also fairly common. Estimates of misdiagnosis as unipolar depression, seems to range from about 40 to 60% of the cases (Ghaemi, Boiman, & Goodwin, 2000; Ghaemi, Sachs, Chiou, Pandurangi, & Goodwin, 1999; Hirschfeld et al., 2003). The fact that depression, and not mania, is the first episode of illness for most patients (Goodwin & Jamison, 2007), can merely offer a partial explanation for the high rate of misdiagnosis. This is supported by a recent study comparing diagnoses given by clinicians at hospital admission, with diagnoses given at the same time by an independent expert psychologist (Øiesvold et al., submitted), which also confirms the high rate of misdiagnosis. The expert performed a structured diagnostic assessment using the Mini International Neuropsychiatric Interview (M.I.N.I.), in conjunction with provided information about patient's symptoms and behaviour from their records. Across the diagnoses present, agreement for bipolar disorders was amongst the poorest, with only poor to fair agreement. While 41 patients were diagnosed with bipolar depression by the expert, only 7 received this diagnosis by the clinicians, and 14 patients were not given a diagnosis of an affective disorder at all by the clinicians.

One explanation for the high rate of misdiagnosis as unipolar depression that has been proposed, is that the introduction of new, safer antidepressants in the 1990s, led to an increase in the diagnosis and treatment of depression overall, and that this may have inadvertently contributed both to increased misdiagnosis and countereffective treatment of bipolar disorder (Ghaemi et al., 1999). However, many reasons for the high rate may to a large extent be the same as the general diagnostic challenges described above, where the most important are patients lack of insight or ability to provide information, clinicians failure to include information from third parties (e.g. family members), clinicians focus on currently presenting symptoms, and attention to euphoric rather than irritable mood in mania. In addition, Goodwin and Jamison (2007) points to the structure of current diagnostic systems as a

contributor to misdiagnosis. For instance, the diagnosis demands a spontaneous manic or hypomanic episode, thus, not taking into account such factors as antidepressant-induced mania in bipolar patients (e.g. Altshuler et al., 1995).

Through the 1970s and 80s, research demonstrated that bipolar disorder (then labelled manic-depressive illness) was often misdiagnosed as schizophrenia, and it was pointed to the symptom overlap between the disorders as an important contributor to this (e.g. Pope & Lipinski, 1978; Taylor, Gaztanage, & Abrams, 1974). The more recent study by Lish and colleagues (1994), reported a rate of misdiagnosis as schizophreniform disorder of 19%, and Hirschfeld and colleagues (2003) reported a rate of 18% misdiagnosed with schizophrenia and 11% misdiagnosed with schizoaffective disorder. This shows that in addition to distinguishing between major affective disorders, distinguishing bipolar disorder from schizophrenia and schizoaffective disorder, is still a diagnostic challenge. While schizoaffective disorder represent the diagnostic category where neither symptoms of schizophrenia or mood disorders can be said to be dominant, there is still substantial symptom overlap between schizophrenia and bipolar disorder (Adler & Strakowski, 2003; Buckley et al., 2004). Furthermore, this is true for both positive and negative symptoms, and whether or not the patient with bipolar disorder presents with psychotic features. For instance, some degree of symptoms like grandiosity, paranoia, acute irritability, hallucinations, thought disorder, disorganized speech, and catatonic-like excitement, can be indicative of both disorders. Negative symptoms like apathy, social withdrawal and lack of affect and energy are associated with both schizophrenia and depression, and the latter may also include psychotic features. In addition, depression is common in schizophrenic patients, both during and after psychosis (Birchwood, Iqbal, Chadwick, & Trower, 2000). In cases where symptom overlap makes the correct diagnosis less clear-cut, it is important to take into account the patients premorbid functioning, family history, course of illness and the nature of any previous episodes (Goodwin & Jamison, 2007). These are all factors that are directly assessed in the DIP, in addition to purely symptomatic factors.

In terms of misdiagnosis of bipolar disorder as a personality disorder, emotionally unstable personality disorder has received the most attention. Emotionally unstable personality disorder is described diagnostically as varying more through the lifespan than the other personality disorders, and there is an ongoing controversy as to whether this disorder may be a part of the bipolar spectrum (Benazzi, 2008; MacKinnon & Pies, 2006). The

overlap between the two disorders, which is especially evident when looking at bipolar II and rapid cycling forms of bipolar disorder, should nevertheless be a smaller challenge with careful attention to liability, reactivity, and the overall symptom cluster (Goodwin & Jamison, 2007). Patients with bipolar disorder are also misdiagnosed with anxiety disorders and substance (including alcohol) abuse disorders. Both these latter groups of disorders are also comorbid with bipolar disorder in about 40% of patients (McElroy et al., 2001). Thus, explanations for the misdiagnosing as these disorders may be that their symptoms have been more prominent than symptoms of bipolar disorder. Though any form of misdiagnosis can have negative consequences for the patient, the most likely pitfall seems to be misdiagnosing bipolar disorder as another affective disorder (unipolar depression), schizoaffective disorder, or a non-affective psychotic disorder (e.g. schizophrenia or schizophreniform disorder).

Consequences of misdiagnosis

Misdiagnosis and delayed diagnosis of bipolar disorder may have detrimental consequences for the progression and prognosis of the disorder. An illustration of this is found in a study by Awad, Rajagopalan, Bolge, and McDonnell (2007), who examined quality of life in bipolar patients misdiagnosed with major depressive disorder. While major depressive disorder and bipolar disorder are in themselves recognized as having a great impact on quality of life, misdiagnosed patients had an even poorer quality of life than the correctly diagnosed patients. Some commonly known negative consequences of delayed treatment of bipolar disorder, in terms of psychosocial implications and possible disease-related neuroanatomical damage are outlined by Berk and colleagues (2007), who also point to the potential neuroprotection that may be provided by correct medication. Delayed treatment is also linked to increased risk of comorbidities (especially substance abuse disorders), forensic complications as a result of committing felonies during illness episodes, and impairments in age-specific developmental tasks. As mentioned earlier, there is also a considerably higher risk of suicide among patients with bipolar disorder, as compared to the general population (e.g. Mork et al., 2009; Tondo et al., 2003; Ösby et al., 2001).

Negative consequences of misdiagnosis as depressive disorder may be particularly severe, and is in part related to the use of psychopharmaceuticals. The results of a study conducted by Matza, Rajagopalan, Thompson, and Lissovoy (2005) indicates treatment patterns for misdiagnosed patients with bipolar disorder, and shows that it differs from the

treatment received by patients correctly diagnosed with both major depression and bipolar disorder. While antidepressants are an effective component in treatment of depressive disorder, they can have a negative effect on the outcome for patients with bipolar disorder. First, it has been shown that the use of antidepressants in bipolar disorder can cause a switch from a depressive to a manic state, and second, it can cause the disorder to progress in direction of more rapid cycling of depressive and manic states (Altshuler et al., 1995; Ghaemi et al., 1999; Ghaemi et al., 2000). Patients with bipolar disorder treated for depression with antidepressants alone, and patients in which initiation of mood stabilizers is delayed at illness onset, may also in general be at a higher risk of hospitalization and suicide, have poorer social adjustment, and represent higher health care costs (Goldberg & Ernst, 2002; Shi, Thiebaud, & McCombs, 2004). In addition, allowing bipolar disorder to progress untreated may have negative consequences because some mood stabilizers seem to become less effective the more affective episodes a patient has experienced (e.g. Swann, Bowden, Calabrese, Dilsaver, & Morris, 1999).

Taken together, the high rates of misdiagnosis and delayed diagnosis, and the detrimental consequences of delayed, ineffective or harmful treatment, leaves no doubt that there is great need of better methods and routines when diagnosing bipolar disorder. There has been a growing awareness of the need for more accurate diagnosis of several psychiatric disorders the recent years, and this has driven the development and research on a number of more or less structured diagnostic interviews.

Diagnostic interviews

The low prevalence of bipolar disorder makes it difficult to study. Nevertheless, the development and testing of diagnostic interviews for use both in research and in clinical practice, continues to contribute to better identification and measurement of specific disorders. These range from fully structured interviews that leave little room for clinical interpretation, to more extensive semi-structured interviews that opens up for clinical interpretation. In general, it has been shown that the use of operational criteria (e.g. Sartorius et al., 1993), and corresponding diagnostic interviews (e.g. Miller, Dasher, Collins, Griffiths, & Brown, 2001; Rogers, 2003; Spitzer, Williams, Gibbon, & First, 1992), has led to more valid and reliable psychiatric diagnoses.

Structured interviews consists of standardized questions with optional probes and can be performed by trained non-clinicians. Composite International Diagnostic Interview (CIDI), developed by the World Health Organization (WHO), is compatible with both the ICD-10 and the DSM-IV and takes from 15 to 90 minutes to administer. CIDI has been shown to be a reliable assessment tool (Andrews, Peters, Guzman, & Bird, 1995), though variable results have been reported concerning validity, ranging from poor for manic and bipolar disorders (Quintana, Gastal, Jorge, Miranda, & Andreoli, 2007) to excellent for bipolar I (Kessler et al., 2006). Another structured interview is the Mini International Neuropsychiatric Interview (M.I.N.I.), which is designed to measure Axis I disorders of the DSM-IV. Diagnostic information obtained with M.I.N.I. is also compatible with diagnoses in the ICD-10. The interview takes about 45-60 minutes to complete (Sheehan et al., 1998). The M.I.N.I. is thought to be a valid and reliable assessment tool (Lecrubier et al., 1997), but this interview does not go into much detail when assessing symptoms of mania and psychosis.

Semi-structured interviews allow the interviewer to freely ask questions when needed, in addition to the standardized questions and probes. This requires clinical experience when administering the interview. Structured Clinical Interview for DSM Axis I (SCID I) is a semi-structured interview compatible with axis I disorders in DSM-III-R, that takes 1 to 2 hours to complete. Skre, Onstad, Torgersen, and Kringlen (1991) assessed the inter-rater reliability of the Norwegian version of SCID I for DSM-III, and found that it yielded highly reliable diagnoses. SCID I covers both affective and psychotic symptoms, and is widely used when diagnosing such disorders. Schedules for Clinical Assessment in Neuropsychiatry (SCAN), developed by the WHO, is another semi-structured interview. The interview takes from 15 to 90 minutes to administer and is compatible with the ICD-10. SCAN 2.1 is thought to be reliable for the assessment of psychiatric disorders (Rijnders et al., 2000).

Semi-structured interviews usually gives the most complete clinical picture and are therefore thought to provide more valid assessments (Brugha, Jenkins, Taub, Meltzer, & Bebbington, 2001). On the other hand, inter-rater reliability may be reduced compared to fully structured interviews, due to potential differences in the clinician's interpretation of the reported symptoms. In general, the most commonly used diagnostic interviews have been shown to have at least acceptable validity and reliability in their original form, though they may be insufficient for the thorough diagnostic assessment of low prevalent severe mental disorders, such as bipolar disorder. While the DIP aims to be more specific in assessing the

psychotic and affective symptoms of low prevalent disorders, many other diagnostic interviews have a more general approach to psychiatric symptoms and syndromes, and may be more suitable as screening instruments when faced with the most severe mental disorders. In addition, despite their widespread use in Norwegian psychiatric health care, there seems to be relatively few studies assessing validity and reliability for the interviews' translated versions. As for international validations of the various assessment tools, there are some published studies validating CIDI (Cho et al., 2002; Quintana et al., 2007), M.I.N.I. (Kadri et al., 2005) and SCAN (Cheng et al., 2001; Hesselbrock, Easton, Bucholz, Schuckit, & Hesselbrock, 1999; Rogers, 2001), and the results of these have been variable. The fact that there are few published studies on the validity and reliability of the numerous translations of diagnostic interviews, and the variable results found in the ones conducted, highlight the importance of such studies when applying new diagnostic tools in translated versions.

To conclude, there are a number of diagnostic interviews being used to measure psychiatric disorders, but few measuring low prevalent, severe psychiatric disorders specifically, and in a satisfactory manner. For example, while SCID I covers both affective and psychotic disorders, and has shown to be both valid and reliable, its user-friendliness has been questioned because of a cumbersome layout and long, complicated questions. On the other hand, other interviews may be easier to administer, but are insufficient for the thorough assessment of the most severe mental disorders. Thus, clinicians may be more inclined to use user-friendly interviews, even though they are considered less useful when diagnosing severe affective and psychotic disorders. However, the main problems with existing diagnostic interviews, are that they are syndrome oriented, and not symptom oriented. Thus, diagnoses derived from a SCID, SCAN, or M.I.N.I. interview are fixed to the diagnostic systems inherent in the interviews, and are not easily converted as diagnostic definitions change, as they are about to with the arrival of DSM-V and ICD-11. In this regard, the DIP differs from other diagnostic interviews by its independence from current diagnostic categories and definitions.

The Diagnostic Interview for Psychoses (DIP)

The Diagnostic Interview for Psychoses (DIP) (Castle et al., 2006) was originally designed for the Australian National Mental Health Survey – low prevalent (psychotic) disorders study, conducted in 1997-1998. The DIP is a semi-structured interview that aims to

bridge the gap between fully structured and lay interviews, and is meant to be used by interviewers with clinical experience. The structure of the interview allows the clinician to make diagnoses based on the same procedure each time, and at the same time opens up for the clinician's use of knowledge and experience during the interview. The complete DIP covers the following main areas: (a) demographic data, (b) social functioning and disability, (c) symptoms, signs and past history items required for the diagnosis of psychotic disorders (diagnostic module, DIP-DM), and (d) patterns of service utilization and patient-perceived unmet needs for services. The DIP-DM can be conducted alone, and takes about 60-90 minutes to complete.

The interview opens up for reliability testing on the level of symptoms, diagnosis and across different theory driven clusters of symptoms. The DIP-DM (see appendix A for an overview) consist of a number of questions that is based on the OPCRIT 90-items checklist, version 3.31 (McGuffin, Farmer, & Harvey, 1991; Williams et al., 1996). The main and follow-up questions of the DIP-DM are either made especially for this interview, or have been adopted from SCAN. The purpose of the questions is to obtain information about the presence of single symptoms and signs covered by the OPCRIT. The questions are formulated to allow the interviewer to ask about present state, past year, and lifetime prevalence of symptoms and signs. They are also arranged in a way that creates a natural progression in the interview, and there are separate sections for areas such as depressive and manic symptoms, psychotic traits, use of drugs, and duration and course of the disorder. In addition, symptoms of affective and psychotic disorders are thought to be identified more precisely when using the DIP, because it has a more detailed level of assessment of many symptoms and illness factors, than other diagnostic interviews. As an example, the DIP operates with separate items for reduced and increased appetite, while the M.I.N.I. and the SCID I operates with only one item assessing changed appetite.

The DIP allows for use of information from other sources, in addition to the information that is obtained during the interview. When necessary, the use of other sources, such as hospital case notes and other informants, is encouraged. For instance, it may be useful to interview a family member to get extended information about premorbid functioning or family history of psychiatric illnesses. Based on observations during the interview, or reports from third parties, the presence of symptoms manifested in behaviour, such as inappropriate affect, agitated activity and pressure of speech, is registered. The responses to each item are

registered in DIP-DM4, the computer software that yields diagnoses according to all known diagnostic systems based on the OPCRIT-algorithm. Thus, the DIP allows for information to be obtained both on the level of individual symptoms and illness factors, and on the level of clusters of symptoms as defined in the current diagnostic classification systems.

Aim of the study

This is the first study of the Norwegian translation of the DIP. The aim of the study is to assess the reliability and applications of the Norwegian version of the DIP, with emphasis on its usefulness in diagnosing bipolar disorder. The results reported here are based on data derived from a larger, ongoing research project assessing the reliability and validity of the DIP. As such, the results are based on a relatively small and preliminary sample, and should be viewed only as indicative of what the final results of the research project may show.

Method

Translation procedure

The diagnostic module of the Diagnostic Interview for Psychosis (DIP-DM) was translated to Norwegian on the basis of the original English language version of November 2008 (University of Western Australia), by translators with experience with both Norwegian and English psychiatric terminology, whose native language was Norwegian (see appendix D for a sample of the Norwegian translation of DIP). Validation of the translation was made by back-translation by an experienced bilingual psychiatrist. To ensure content validity, inconsistencies and related problems were discussed in the research team, and the Norwegian translation was continuously corrected accordingly. The translation procedure is still an ongoing process as part of the larger research project that the study reported here is derived from.

Sampling

It is important that reliability studies of assessment tools are conducted in populations and settings where factors that may influence the diagnostic process, is similar to those found in populations and settings where the assessment technique will ultimately be applied (Thompson & Walter, 1988). For the purpose of this study, this means populations with high prevalence of symptoms of psychotic and affective disorders who are in contact with Mental

Health services. Accordingly, the respondents were drawn from psychiatric hospital wards at The University Hospital for Northern Norway in Tromsø and Nordland Hospital in Bodø. The study was approved by the Regional Committees for Medical and Health Research Ethics (REK).

Patients under the age of 18, with language difficulties, dementia or mental retardation were excluded. Two of the cases in the original sample were excluded from the data analysis, one because of insufficient responses to interview questions, and one because of withdrawal from participation in the study. In total, 27 patients were included in the *inter-rater reliability* study, of which 14 were interviewed on a second occasion and included in the assessment of *test-retest reliability*. The first interview was conducted as part of the patient's global diagnostic assessment, after which he or she was asked for written consent for the data to be included in the research project (see appendix E for the written consent form and its accompanying information). This was done by a member of the research staff, who gave a full explanation of the purpose of the study, provided the patients with the details of the study in writing, informed them that participation was voluntary, and that they could withdraw from the study at any time without this affecting their clinical care. In addition, patients were at this point asked to participate further by allowing a second interview by another interviewer. The patients were given a minimum of 24 hours to consider the request, after which they informed the hospital staff of their decision. A compensation was given to patients who participated with a second interview, by a choice between NOK 250 or 10 tickets in the National lottery.

Data collection

Each patient was first assessed independently with the DIP by two raters from the research team (*inter-rater reliability* test), one of whom conducted the interview, while the other scored from observation. At the end of the session the observer was allowed to repeat any interview questions if there was disagreement with the interviewer about the sufficiency of the information that had been obtained, or about the use of skips between and within sections (the interview contains a number of built-in cut-offs and skips between sections to avoid redundant questioning when initial screening questions have indicated that psychopathology is unlikely to be present in that section). This helped ensure the same potential for variability between raters, as would have been present had the interviewer and

observer not assessed the patient in the same instance. In four cases, combined sources, using case notes in addition to the semi-structured interview with the respondent, was used. After the initial rating, a third interviewer, blind to the results of the first interview, made an independent assessment of 14 of the patients (*test-retest reliability* test).

Raters and training

The assessments were conducted by two pre-graduate students of clinical psychology in training, one clinical psychologist, two psychiatrists and two resident doctors. The seven raters were assigned the roles of interviewer, observer, and re-interviewer. The majority of the raters (five raters) were trained by the principal researcher from the Italian study of DIP, Rossi. He, in turn, had been trained by Jablensky in Australia. Two more raters were trained by the Norwegian interviewers. A balanced design, in terms of rotation of raters in the roles of interviewer, observer, and re-interviewer, was strived for. On average, the raters assumed the role of interviewer 3.85 times, (range 1-8), observer 3.85 times, (range 1-6), and re-interviewer 2.14 times, (range 1-5).

Data analysis

In the *inter-rater reliability* analysis, every DIP rated by the interviewer and the observer was compared, and in the *test-retest reliability* analysis, the interviews rated by the interviewers of the first and second interview were compared. The *kappa statistic* (Cohen, 1960) was used for ratings on a dichotomous scale, and the *intraclass correlation coefficient (ICC)* (Shrout & Fleiss, 1979) was used for ratings on an ordinal scale. For each pair of interviews, all the 90 individual items in the Operational Criteria for Psychosis (OPCRIT) and the ICD-10 and DSM-IV diagnoses were considered in the analysis. Three items, source of information, time frame, and sex code, were excluded from the summaries of results provided in this paper, as they are not of diagnostic value and unlikely to be disagreed upon. One OPCRIT item, relationship between psychotic and affective symptoms, is coded on a nominal scale with four alternatives. For the analysis, this item was re-coded and reliability indices calculated separately for each of the four alternatives. With exclusion of three items, and with the addition of the re-coded variables, the total number of individual items assessed, is 90. On the diagnostic level, the frequency of each of the specific diagnoses was insufficient for the use of the *kappa statistic* in some cases. For this reason, the *kappa* for broader

diagnostic categories relevant to distinguishing bipolar and manic disorders from other disorders, are reported. The categories used were bipolar/manic disorders, depressive disorders and non-affective psychoses (including schizoaffective disorder). As described earlier, the differentiation of bipolar and manic disorders from these groups of disorders is diagnostically challenging. Thus, analysis at this level of categories can be considered highly relevant to the purpose of this study; assessing the usefulness of the DIP in diagnosing bipolar disorder.

In both the *inter-rater* and *test-retest* comparisons, the *overall pairwise agreement* (PAR) and *kappa statistic* was used to measure agreement between raters on individual dichotomous items, and on specific and broad diagnostic levels. The *kappa statistic*, though often the preferred index of diagnostic agreement in psychiatric research, as it adjusts the observed rate of agreements for agreements due to chance, has the disadvantage of being affected by the prevalence of the symptom or disorder. This means that items that show high sensitivity and specificity may have low predictive accuracy if the prevalence of the symptom or the disorder is low (Feinstein & Cicchetti, 1990). Cicchetti and Feinstein (1990) proposed a solution to this problem, whereby the kappa index should always be accompanied by the observed proportions of 'positive' agreement, p_{pos} (i.e. agreement on the presence of the symptom) and 'negative' agreement, p_{neg} (i.e. agreement on the absence of the symptom). Accordingly, these values were calculated in addition to all *kappa* values. Because many items in the DIP assessing specific symptoms and illness factors are rated on an ordinal scale, we chose to calculate the *ICC* for these particular items, as a measure of agreement between raters. *ICC* takes into account the relative distance between the rated values, and as such it can be considered a more accurate assessment of agreement, than the alternative of recoding items into dichotomous variables for the *kappa statistic*. For example, a case where one rater scored a symptom as present for a month, and the other rater as present for two weeks, would contribute to a higher *ICC* value, than if the second rater had scored it as not present (see appendix F for examples of items rated on a dichotomous and an ordinal scale). In addition, *ICC* has been shown to be mathematically equivalent to the alternative *weighted kappa* (Fleiss & Cohen, 1973). As such, it is chosen as the measure of reliability for OPCRIT items with ordered response categories, due to its more manageable and straightforward use, and is interpreted along with *kappa* in terms of Landis and Kochs (1977) proposed standards for interpretation of *kappa* values, where <0 signifies no agreement, 0-0.19 poor agreement,

0.20-0.39 fair agreement, 0.40-0.59 moderate agreement, 0.60-0.79 good agreement, and 0.80-1.00 excellent agreement.

SPSS 16.0 (SPSS Inc., 2007) was used for data management, descriptive statistics, Cohens *kappa* and *intraclass correlation*. For assessment of overall pairwise, positive and negative agreement, the DAG_Stat spreadsheet (MacKinnon, 2000) was used.

Results

Reliability study sample

27 respondents were included in the *inter-rater reliability* study, and 14 of these were interviewed again for the *test-retest* study. The mean age of the respondents was 38.5 years (SD=16.59). 15 of the respondents (55.6%) were female, and 12 (44.4%) were male. Socio-demographic characteristics are assessed in the DIP. Nine of the respondents (33.3%) had been married or living with a partner for more than six months at some time in life, and four of the respondents (14.8%) had been unemployed at illness onset.

The mean time between the first (*inter-rater*) and second (*retest*) interviews was 19.5 days, and the median was 7 days between each interview (range=2-90). The time spent on each interview was approximately 75 minutes (range=40-120) in both the *inter-rater* study and the *test-retest* study.

Inter-rater reliability

Table 1 shows the *ICC* and *kappa* values with their respective 95% confidence intervals, for a selection of the 90 items of the Operational Criteria for Psychosis (OPCRIT) checklist, consisting mainly of items for affective symptoms and items considered especially relevant for the purpose of this study. Also included in the table, is the *overall pairwise agreement* (PAR), and positive (p_{pos}) and negative (p_{neg}) agreement, for items where kappa was calculated. *Inter-rater* reliability was generally high. Using the *kappa statistic* and the *intraclass correlation coefficient (ICC)*, 70% (63) of the items achieved a value of ≥ 0.60 , i.e. good to excellent agreement, with 26.7% (24) of all the items in the >0.80 range. Values in the range of 0.40-0.59, i.e. moderate agreement, was obtained for 12.2% (11) of the items, while a value of 0.39, i.e. fair agreement, was obtained for 1.1% (1). For the remaining 16.7% (15) of the items, results were not significant (11 items), or reliability indices could not be calculated due to frequency of responses in one category of dichotomous ratings being too

Table 1. *Inter-rater Reliability for Selected OPCRIT Items*

	<i>n</i> ^a	<i>p</i> _{pos}	<i>p</i> _{neg}	Overall pairwise agreement	Kappa ^b	ICC ^c	95% CI	Level of agreement
Mode of onset	27	-	-	-	-	.80*	.61-.91	Excellent
Psychosocial stressor prior to first episode	17	.83	.80	.82	.63*	-	.34-.92	Good
Poor premorbid work adjustment	10	.89	.94	.93	.83*	-	.61-1.06	Excellent
Family history of psychiatric disorder other than schizophrenia	20	.95	.88	.93	.82*	-	.59-1.06	Excellent
Family history of schizophrenia	5	.98	.98	.98	.87*	-	.86-1.05	Excellent
Dysphoria	20	-	-	-	-	.95*	.89-.98	Excellent
Loss of pleasure	19	-	-	-	-	.86*	.71-.94	Excellent
Suicidal ideation	22	-	-	-	-	.82*	.65-.91	Excellent
Poor concentration	18	-	-	-	-	.61*	.31-.80	Good
Slowed activity	10	-	-	-	-	.81*	.63-.91	Excellent
Loss of energy	17	-	-	-	-	.92*	.83-.96	Excellent
Changed libido	11	.62	.86	.80	.57**	-	.09-.87	Moderate
Initial insomnia	13	-	-	-	-	.60*	.29-.79	Good
Middle insomnia	13	.87	.90	.89	.77*	-	.53-1.01	Good
Early morning wakening	6	-	-	-	-	.94*	.87-.97	Excellent
Excessive sleep	12	-	-	-	-	.78*	.57-.89	Good
Excessive reproach	13	-	-	-	-	.74*	.50-.87	Good
Delusions of guilt	8	-	-	-	-	.70*	.43-.85	Good
Elevated mood	12	-	-	-	-	.93*	.84-.97	Excellent
Irritable mood	14	-	-	-	-	.70*	.45-.85	Good
Thoughts racing	14	-	-	-	-	.70*	.42-.85	Good
Distractibility	13	-	-	-	-	.85*	.71-.93	Excellent
Excessive activity	13	-	-	-	-	.81*	.62-.91	Excellent
Reduced need for sleep	19	-	-	-	-	.61*	.31-.80	Good
Reckless activity	11	-	-	-	-	.70*	.44-.85	Good
Increased sociability	9	-	-	-	-	.54**	.21-.76	Moderate
Increased self-esteem	12	-	-	-	-	.78*	.58-.90	Good
Hallucinations in any modality	20	-	-	-	-	.79*	.59-.90	Good
Grandiose delusions	9	-	-	-	-	.58*	.27-.78	Moderate
Bizarre delusions	12	-	-	-	-	.69*	.43-.85	Good
Lack of insight	8	.93	.97	.96	.91*	-	.73-1.08	Excellent
Impairment/incapacity during disorder	27	-	-	-	-	.66*	.38-.83	Good
Psychotic and affective symptoms never together	15	.33	.86	.76	.71*	-	-.31-.73	Good
Non-affective psychotic symptoms dominates	7	.60	.91	.85	.51**	-	.10-.93	Moderate
Non-affective psychotic and affective symptoms balanced	4	-	-	-	ns	-	-	-
Affective symptoms dominates	8	.77	.93	.89	.70*	-	.40-1.00	Good
Agitated activity	9	-	-	-	-	.58*	.27-.78	Moderate
Bizarre behaviour	5	.57	.94	.89	.52**	-	.07-.97	Moderate
Inappropriate affect	1	-	-	-	-	.80*	.61-.90	Excellent
Pressure of speech	12	-	-	-	-	.56*	.25-.77	Moderate

Note. CI = Confidence Interval

^a Item rated as present by any rater. Where ICC is calculated, as present for any period of time by any rater. ^b ICC not calculated for dichotomous variables. ^c Kappa not calculated for ordered variables.

p*<.001. *p*<.01. ****p*<.05. ns = not significant.

low (4 items). For the 24 items where *kappa* was calculated, 15 p_{pos} and 22 p_{neg} values were 80 or better, indicating a moderate positive agreement and a high negative agreement. PAR was above 0.80 for all these 24 comparisons, except one which was 0.76.

Table 2 shows the overall PAR, p_{pos} and p_{neg} values, and *kappa* values with their 95% confidence intervals, for the broad diagnostic categories according to both ICD-10 and DSM-IV. At this level, agreement between raters was also high. Within both diagnostic systems, *kappa* values ranged from 0.63 to 0.76 for the categories of depressive disorders and non-affective psychoses, indicating good agreement, while *kappa* was 0.82 for the category of bipolar/manic disorders, indicating excellent agreement. PAR was above 0.80 for all comparisons, indicating good agreement beyond chance. Values of p_{pos} and p_{neg} were all above 0.80, with negative agreement slightly better than positive. On a specific diagnostic level, *kappa* was not calculated if the diagnosis had been rated as present by any rater in less than four cases, and thus, was calculated only for five diagnoses. One of these did not reach significance (depression with psychosis within DSM-IV). Agreement on the remaining four diagnoses was good, with a *kappa* of 0.62 ($p=.001$) for schizophrenia and 0.79 ($p<.001$) for bipolar disorder within ICD-10, and 0.65 ($p=.001$) for schizophrenia and 0.66 ($p<.001$) for bipolar II disorder within DSM-IV. PAR, p_{pos} , and p_{neg} values were all above 0.80 for these four specific diagnoses, and negative agreement higher than positive.

Table 2. *Inter-rater Reliability for Broad Diagnostic Categories*

	ICD-10							DSM-IV						
	n ^a	p_{pos}	p_{neg}	Overall pairwise agreement	Kappa	95% CI	Level of agreement	n ^a	p_{pos}	p_{neg}	Overall pairwise agreement	Kappa	95% CI	Level of agreement
Bipolar/manic disorders	9	.88	.95	.93	.82*	.59-1.06	Excellent	9	.88	.95	.93	.82*	.59-1.06	Excellent
Depressive disorders	6	.80	.95	.93	.76*	.43-1.08	Good	6	.80	.95	.93	.76*	.43-1.08	Good
Non-affective psychoses	16	.86	.85	.85	.70*	.44-.97	Good	16	.81	.81	.81	.63*	.35-.92	Good

Note. CI = Confidence Interval

^a Number of diagnoses present in the sample as rated by any rater
 * $p<.001$

Test-retest reliability

Table 3 shows the results of the *test-retest reliability* analysis, for the same selected OPCRIT items as presented in Table 1. Out of the total of 90 items, 3.3% (3) achieved excellent agreement with values above 0.80, 12.2% (11) achieved good agreement with

Table 3. Test-retest Reliability for Selected OPCRIT Items

Item	n ^a	p _{pos}	p _{neg}	Overall pairwise agreement	Kappa ^b	ICC ^c	95% CI	Level of agreement
Mode of onset	14	-	-	-	-	ns	-	-
Psychosocial stressor prior to first episode	9	-	-	-	ns	-	-	-
Poor premorbid work adjustment	4	-	-	-	ns	-	-	-
Family history of psychiatric disorder other than schizophrenia	9	.62	.67	.64	ns	-	-	-
Family history of schizophrenia	3	.80	.96	.93	.76**	-	.32-1.20	Good
Dysphoria	12	-	-	-	-	.56***	.06-.84	Moderate
Loss of pleasure	12	-	-	-	-	.58***	.08-.84	Moderate
Suicidal ideation	11	-	-	-	-	.70**	.27-.89	Good
Poor concentration	8	-	-	-	-	.56***	.05-.83	Moderate
Slowed activity	5	-	-	-	-	ns	-	-
Loss of energy	11	-	-	-	-	.64**	.17-.87	Good
Changed libido	5	.75	.90	.86	.66**	-	.25-1.07	Good
Initial insomnia	7	-	-	-	-	ns	-	-
Middle insomnia	8	.77	.80	.79	.57***	-	.15-1.00	Moderate
Early morning wakening	6	-	-	-	-	ns	-	-
Excessive sleep	7	-	-	-	-	ns	-	-
Excessive reproach	10	-	-	-	-	ns	-	-
Delusions of guilt	4	-	-	-	-	-	-	-
Elevated mood	7	-	-	-	-	.67**	.23-.88	Good
Irritable mood	7	-	-	-	-	ns	-	-
Thoughts racing	11	-	-	-	-	ns	-	-
Distractibility	8	-	-	-	-	ns	-	-
Excessive activity	6	-	-	-	-	.64**	.17-.87	Good
Reduced need for sleep	11	-	-	-	-	ns	-	-
Reckless activity	6	-	-	-	-	.60***	.11-.85	Good
Increased sociability	4	-	-	-	-	ns	-	-
Increased self-esteem	7	-	-	-	-	.63**	.20-.86	Good
Hallucinations in any modality	10	-	-	-	-	ns	-	-
Grandiose delusions	5	-	-	-	-	.57***	.06-.84	Moderate
Bizarre delusions	7	-	-	-	-	.49***	-.06-.80	Moderate
Lack of insight	5	-	-	-	.43 ns	-	-	-
Impairment/incapacity during disorder	14	-	-	-	-	ns	-	-
Psychotic and affective symptoms never together	9	.80	.77	.79	.57***	-	.15-1.00	Moderate
Non-affective psychotic symptoms dominates	5	-	-	-	ns	-	-	-
Non-affective psychotic and affective symptoms balanced	2	-	-	-	ns	-	-	-
Affective symptoms dominates	3	1	1	1	1*	-	1-1	Excellent
Agitated activity	5	-	-	-	-	-	-	-
Bizarre behaviour	3	-	-	-	-	-	-	-
Inappropriate affect	1	-	-	-	-	-	-	-
Pressure of speech	5	-	-	-	-	.44***	-.03-.77	Moderate

Note. CI = Confidence Interval

^a Item rated as present by any rater. Where ICC is calculated, as present for any period of time by any rater. ^b ICC not calculated for dichotomous variables. ^c Kappa not calculated for ordered variables.

p*<.001. *p*<.01. ****p*<.05. ns = not significant.

values in the range of 0.60-.079, and 16.7% (15) achieved moderate agreement with *kappa* or *ICC* values in the range of 0.40-0.59. For the remaining 67.8% (61) of the items reliability estimates were either not significant (39 items), or could not be calculated due to frequency of responses in one category of dichotomous ratings being too low or ordinal ratings showing zero variance (22 items). For the 10 items where *kappas* were obtainable and significant, 5 p_{pos} and 9 p_{neg} values were above 0.80, indicating fair positive agreement and a high negative agreement. PAR was above 0.80 for 7 of these 10 comparisons.

Results of the *test-retest* analysis on a broad diagnostic level can be seen in table 4. At this level, agreement between raters is moderate to high. However, given the low *n* (14) in the *test-retest* study, these results should be viewed with some caution. There was moderate agreement for depressive disorders as classified according to ICD-10 diagnoses, with a *kappa* of 0.58. The categories of bipolar/manic disorders and non-affective psychoses within ICD-10, and depressive disorders and non-affective psychoses within DSM-IV, showed good agreement with *kappa* values ranging from 0.65 to 0.76. Bipolar/manic disorders within DSM-IV achieved perfect agreement with a *kappa* of 1.00. PAR was 0.86 or above for all comparisons, indicating good agreement beyond chance. Values of p_{pos} and p_{neg} were 0.67 or better, and negative agreement was slightly higher than positive. On a specific diagnostic level, *kappa* was not calculated if the diagnosis had been rated as present by any rater, in less than four cases. The four specific diagnoses *kappa* was obtainable for in the *test-retest* analysis, were the same as those that reached significance in the *inter-rater* analysis. The *kappa* for schizophrenia within both ICD-10 and DSM-IV did not reach significance in the test-retest analysis. Agreement for bipolar disorder within ICD-10 was moderate, with a

Table 4. Test-retest *Reliability for Broad Diagnostic Categories*

	ICD-10							DSM-IV						
	^a n	p_{pos}	p_{neg}	Overall pairwise agreement	Kappa	95% CI	Level of agreement	^a n	p_{pos}	p_{neg}	Overall pairwise agreement	Kappa	95% CI	Level of agreement
Bipolar/manic disorders	5	.75	.90	.86	.65***	.21-1.09	Good	5	1	1	1	1*	1-1	Excellent
Depressive disorders	4	.67	.91	.86	.58***	.05-1.10	Moderate	3	.80	.96	.93	.76**	.32-1.20	Good
Non-affective psychoses	8	.86	.86	.86	.71**	.35-1.08	Good	7	.83	.88	.86	.71**	.36-1.07	Good

Note. CI = Confidence Interval

^a Number of diagnoses present in the sample as rated by any rater
* $p < .001$

kappa of 0.58 ($p < .05$). PAR was 0.86, p_{pos} 0.67 and p_{neg} 0.91. Agreement on bipolar II disorder within DSM-IV was perfect with a *kappa* of 1.00 ($p < .001$).

Discussion

The purpose of this study was to test the reliability of the Norwegian translation of the Diagnostic Interview for Psychoses (DIP), with a focus on bipolar disorder. The use of this semi-structured diagnostic interview for severe mental disorders may prove useful in making more accurate diagnoses of bipolar disorder, by ensuring assessment of previous illness episodes and associated illness factors. The results reported here are preliminary, and the data was derived from a larger research project which is the first to assess reliability and validity of the DIP in Norway. In general, the results of this study indicates that the DIP has good reliability, and the findings are consistent with the results of the study of the original Australian version of the DIP (Castle et al., 2006), and a study of an Italian translation (Rossi et al., 2010).

Broad diagnostic agreement

On a broad diagnostic level, there was good to excellent inter-rater reliability, and moderate to good test-retest reliability, both when applying ICD-10 and DSM-IV criteria (see Table 2 and 4). At this level, results indicate that the DIP reliably distinguishes bipolar disorder from other major affective disorders, from non-affective psychoses (including schizophrenia), and from schizoaffective disorder, using the current diagnostic criteria. As described earlier, the delineation between bipolar and these other diagnostic groups has proved to be challenging. This delineation is highly relevant for the correct diagnosis of bipolar disorder, because of the implications for prognosis and treatment. In terms of the somewhat poorer test-retest reliability, than inter-rater reliability, an observed tendency was that the re-interviewer generally recorded a less severe symptomatology. Since only lifetime symptoms were rated, and the re-interview was performed after a relatively short time, this can probably not be attributed only to an improvement in the respondents condition between the two interviews. However, maybe as a result of the short time between the interviews, the respondents may have learned to avoid probes, and thus, under-reported symptoms. Respondents may also have attempted to be both overly consistent, or to do the opposite by offering novel information to avoid repetitiveness (Robins, 1985). An alternative explanation

may also be *attenuation*, a phenomenon in which the respondents show a tendency to report less symptomatology on successive interviews (Jensen et al., 1995).

Agreement on specific diagnoses

Reliability estimates could not be calculated for all specific diagnoses, due to the low frequency of some diagnoses in the sample. However, where estimates could be obtained, inter-rater agreement was good and test-retest agreement moderate or excellent. When taking a closer look at the diagnostic disagreements that were present where at least one of the raters came out with a diagnosis of a bipolar or manic disorder, the specific types of disagreements reflected the difficulties often encountered in clinical practice. In the three cases with such disagreement, the specific diagnoses obtained other than bipolar disorder, were schizophrenia, schizoaffective disorder, and depressive disorder. Depressive disorder, however, was never confused with schizoaffective disorder in this study.

The observed pattern of disagreement confirms previously mentioned issues in differential diagnosis between these disorders. However, where diagnostic disagreements about bipolar disorder are present, the majority of diagnoses given across raters and diagnostic systems are still within the group of bipolar disorders. In addition, in the case where a diagnosis of schizoaffective disorder was given, it was specified as mixed type, which requires that criteria for both schizoaffective disorder and for mixed bipolar disorder are met. Thus, given the overlapping symptoms of these disorders, and the disputed boundaries between them, the observed disagreements were not surprising. As mentioned earlier, the validity of a clear distinction between bipolar disorders, schizoaffective disorders, and schizophrenic disorders, has been disputed. Accordingly, the observed disagreements on the diagnostic level, may be due both to insufficiencies in the DIP itself, in the scoring made by raters, or in the criteria of the diagnostic systems that the DIP generates diagnoses according to. It may be that other illness criteria or dimensions than the ones currently used, need to be considered, or that the relative importance given each of the existing criteria needs to be changed. Nevertheless, this shows the need for further research on how to define and differentiate these disorders, as such knowledge must be the foundation for both diagnostic systems and diagnostic tools.

ICD-10 versus DSM-IV

No analyses were conducted specifically aimed at differences between the two diagnostic systems. However, in addition to disagreements between the raters, a striking tendency of disagreement relevant to bipolar disorder was observed between diagnoses generated according to the ICD-10 and the DSM-IV. Across all raters, there were 19 instances with agreement between the two diagnostic systems on a diagnosis of bipolar disorder or mania (i.e. agreements between the diagnostic systems about the resulting diagnosis derived from one single rater's recordings of one specific patient). However, in 13 of these cases the DSM-IV yielded a diagnosis of a less severe disorder, bipolar II, where the ICD-10 yielded a diagnosis of bipolar affective disorder. This may confirm what has already been discussed - that the DSM-IV has a more strict definition of the disorder than the ICD-10, and it demonstrates how one person can get two different diagnoses depending on the diagnostic system used.

In light of the known clinical issues in detecting bipolar disorder, the DIP has an advantage when diagnosing bipolar disorder in that it yields diagnoses according to several diagnostic systems, including the ICD-10, which may be more inclusive. This feature of the interview is also important given the fact that there is still a lack of consensus about the optimal classification of severe mental disorders (Kendell & Jablensky, 2003; Castle & Jablensky, 2005). Nevertheless, the fact that most of the respondents in this study were inpatients in psychiatric hospital wards, raises concerns about the high rate of bipolar II disorder diagnoses given according to the DSM-IV criteria. It would be expected for bipolar I disorder to be more prevalent than bipolar II disorder in this particular population, because of the severity required to be admitted to a psychiatric hospital. However, none of the respondents received a diagnosis of bipolar I disorder according to the DSM-IV criteria.

The severity of the disorders present in the sample of this study, can be seen when taking a closer look at the two core symptoms of mania assessed in the DIP, which also serve as skip criteria for the module for manic symptoms if they are not present. Out of all 27 respondents included in the study, 10 had experienced elevated mood and 13 irritable mood in the most severe form, as rated by any rater. Specifically, this means that the symptom persisted for at least two weeks or, if it had a shorter duration, led to hospital admission due to its severity. Thus, a professionals' clinical evaluation based on the DSM-IV criteria, would be unlikely to conclude with the diagnosis of bipolar II disorder for these particular patients,

and this leads to suspicion about possible flaws in the computer software used. For this reason, investigations of what specific factors causes this high rate of bipolar II disorder, should be conducted to rule out possible errors in the DIP software or the algorithm it uses.

Agreement on specific symptoms and core items for mood disorders

On the level of individual symptoms and signs (item level), inter-rater reliability ranged from moderate to excellent. When looking specifically at the 29 items assessing symptoms of mood disorders, 10 obtained excellent agreement and 15 good agreement. Two items, changed libido and increased sociability, obtained moderate agreement. Reliability indices could not be calculated for two items. Test-retest reliability for the 14 affective symptoms which had high enough frequency for ICC or kappa to be estimated, ranged from moderate to good, with 7 items obtaining good agreement and 7 moderate agreement. In terms of the two items in the inter-rater reliability study that obtained only moderate agreement, this may be related to the interviewers. A possible explanation for the lower agreement for changed libido may simply be that sexuality is a matter that both respondents and interviewers are reluctant to discuss. Respondents may also have been more open about this to same gender interviewers, or the counter-intuitive manner of scoring on this particular item as opposed to other items, may have resulted in errors on the part of the raters. Changed libido is not scored on a continuum according to duration or severity as most other items, but as not present, increased libido, or diminished libido, and interviewers may not have been completely observant of this deviation when scoring. As for the item increased sociability, there is a possibility that this item is merely more open to subjective interpretation on the part of the raters than other items. Consequently, when training interviewers, special emphasis must given to scoring of single items which stand out from the general rules of the DIP.

In general, the fact that agreement ranged from good to excellent for the majority of items assessing affective and psychotic symptoms (44 out of 51), is promising in establishing the reliability of the DIP as a useful tool for differential diagnosis of severe mental disorders. Of particular interest when looking at agreement for individual items relevant for bipolar disorder, are the core symptoms of depression and mania. The main items in this regard, dysphoria and elevated mood, showed especially good agreement with ICC values above 0.90. The other two items that can be considered to assess the core symptoms of these mood disorders, loss of pleasure and irritable mood, also showed excellent and good agreement,

respectively. When using the DIP, accuracy on these items is very important because the responses that are recorded here decide whether or not to use the built in skip criteria for the modules assessing depression and mania. As mentioned earlier, assessment of the existence and nature of any previous illness episodes is diagnostically very important, but not always thoroughly conducted in clinical practice. The DIP allows for this to be done in a structured manner, while at the same time avoiding redundant questioning by use of skip criteria.

Agreement concerning associated features, course, and severity of illness

The DIP includes direct assessment of non-symptomatic factors associated with different disorders, that may be of relevance to differential diagnosis and treatment. For instance, family history of mental disorders is assessed by two items, distinguishing between family history of mental disorder with and without schizophrenia. In the inter-rater analysis, agreement on these items was excellent, while in the test-retest analysis, only the item for family history with schizophrenia reached significance and there was good agreement on this item. Another potentially important factor is premorbid functioning. This is assessed in terms of premorbid occupational and social functioning separately. Only the estimate for occupational functioning in the inter-rater analysis reached significance. However, agreement on this particular item was excellent. Thus, it may be that with a larger sample size, all these important items would be shown to be reliably assessed with use of the DIP. Agreement on complicating factors such as alcohol and substance abuse (both associated with and independent of psychopathology) were also assessed, and ranged from good to excellent.

An item that is given particular emphasis in the DIP, concerns what type of symptoms are most dominant in the patient's clinical picture (relationship between psychotic and affective symptoms). This item discriminates between disorders that are characterized mainly by affective versus psychotic symptoms, or alternatively where both types of symptoms are balanced. Thus, this item is important for the distinction between affective disorders, schizoaffective disorders, and non-affective psychotic disorders. Specifically, the respondent is asked if affective and psychotic symptoms have been experienced during the same period of time, or during separate illness periods. As can be seen in Tables 1 and 3, only four of the eight possible tests of reliability for the re-coded versions of this item reached significance. However, agreement on these ranged from moderate to good, and this is encouraging given the importance of this item.

The severity of a given disorder is also a dimensional aspect of clinical relevance. In addition to the assessment of severity inherent in the coding of items aimed at specific symptoms, severity is assessed in terms of total duration of the illness, impairment or incapacity during the disorder, and deterioration from premorbid level of functioning. In terms of the duration of the illness, almost all respondents (20 of 27) reported the maximum duration of 99 weeks allowed in the DIP. Thus, a meaningful estimate of reliability could not be obtained for this item. For the items impairment or incapacity during the disorder and deterioration from premorbid level of functioning, the reliability estimate was significant only for the former in the inter-rater analysis, and showed good agreement ($ICC=.66$). For these two items, all respondents but one responded confirmatory both in the first and second interview, and this explains why significant reliability estimates could not be obtained. The reason why reliability could be estimated for one item (impairment or incapacity during the disorder) in the inter-rater analysis, is that this item is scored on an ordinal scale according to severity of impairment (while the other item is scored dichotomously), giving room for more variation in responses. The high rate of positive responses on items assessing the severity of the disorder is not surprising in a population of inpatients on psychiatric hospitals, given the severity of disorder required to be admitted. However, for the purpose of reliability testing of the DIP, it may be useful to investigate the reliability of these items in a population with less severe psychopathology, which would yield more variation in responses. Consequently, the ongoing sampling for the final assessment of the reliability and validity of the DIP now also includes outpatients.

An important factor that has been pointed to as relevant to differential diagnosis, but that may be a source of concern within the DIP, is course of disorder. This item assesses to what degree the illness can be characterised as chronic or as consisting of more or less separate episodes. Only moderate agreement was obtained for this item. A reason for this might be that this item does not have a question specified in the interview. Instead, it is up to the interviewer whether to base their scoring of this item on information obtained previously in the interview, or to ask the respondent directly about the course of their disorder. Accordingly, it might be advisable to improve this item in future revisions of the interview, by making more specific instructions to the interviewer.

In general, the DIP incorporates a number of items aimed at assessing the associated features, course, and severity of severe mental disorders, and the estimates that were

obtainable in this study are promising by indicating that these may be reliably assessed. Not all of these associated factors are included in the current diagnostic systems, in ways that allow for them to be decisive in the generation of diagnoses by use of the computer software that accompanies the DIP. Nevertheless, the information yielded from the questions on these items during the interview, may be very important for the final diagnostic decision. Consequently, the final diagnostic decision should always be conducted by using the information obtained by use of diagnostic tools, such as the DIP, in conjunction with a professional's clinical evaluation.

Limitations

The sample size in this study is in the lower range of what can be considered acceptable for calculating reliability indices, and on the broad diagnostic level, the number of cases falling within each category as rated by any rater, is relatively low. Only 9 cases fell within the category of bipolar disorders, 6 in depressive disorders, and the remaining 16 in non-affective psychoses. Generally, a small sample size can generate low statistical power, and a larger sample size would possibly have yielded more stable results. As can be seen in Tables 1 to 4, the confidence intervals of the reliability estimates are generally wide, and this indicates that results may not be stable. The ongoing sampling for the final assessment of the validity and reliability of the DIP, continues until the sample size is large enough to achieve satisfactory statistical power for these estimates to be made. Thus, as stated earlier, the results reported in this paper must be viewed with caution, and only as an indication of what the final assessment of the Norwegian translation of DIP may show.

Methodological issues may have contributed to the poorer test-retest agreement, as compared to inter-rater agreement. More specifically, insufficient information may have been provided to the respondents before the re-interview, about the purpose of the second interview. Given that the respondents had already completed the interview once, they may have falsely believed that the interviewer conducting the second interview, already had information about them and consequently held back information. In retrospect, clearer specifications of the procedure that was to be followed by the interviewers, and of the instructions that should be given the respondents, could possibly have been used to reduce variations due to circumstantial factors not directly related to the DIP itself.

Other methodological issues may also be raised. First, there are issues concerning the

raters. Differences between them both in terms of training with the DIP and in clinical experience and training, may have influenced the results in a negative manner. However, the fact that reliability was generally good indicates that the DIP can successfully be used by interviewers with different clinical backgrounds, given that they receive adequate training. As can be seen where the raters are described, a fully ideal rotation of raters was not obtained. However, no systematic effects of this were observed. Second, at least one issue is related to the specific population the respondents were drawn from, which consists of patients with severe mental disorders in a more or less acute phase of the disorder. Rapid treatment effects may have resulted in respondents giving a different presentation of their symptomatology in the second interview, and this may be a potential threat to test-retest reliability. The severity of most respondents disorders, may also have contributed to the considerably longer time spent on each interview in this study (approximately 75 minutes), as compared to the time spent on each interview in the Australian and Italian studies (20-40 minutes). The time spent on each interview varies according to the number of symptoms reported by the patients, the specific use of skip criteria this leads to, and whether the interview is used only for measuring lifetime symptoms, or both lifetime and present state symptoms. Because the interviews that were conducted by the research team was also to be used as part of the current diagnostic evaluation for many of the respondents, both lifetime and present state were assessed for these patients. All seven raters found it hard to make a proper diagnostic evaluation of both these time periods using less than 60 minutes. However, the increased time spent on each interview in this study, does not seem to affect the applicability of DIP.

Further research on the DIP

The DIP is a relatively new diagnostic interview, and the assessment of the reliability and validity of the Norwegian translation is still in progress. Furthermore, the original interview is continuously being revised as new experiences are made from its use, and as research on severe mental illness in general progress. Thus, continued effort to establish its reliability, validity, and applicability, is an important area for future research. This also applies to translations of the DIP to other languages, and the application of the DIP in other countries and areas might be useful in terms of cross-cultural and cross-national comparison and cooperation between those working in the area of severe mental disorders. Because bipolar disorder is a low prevalent disorder, work on how to overcome the diagnostic

challenges clinicians experience when faced with it, and further development of the diagnostic tools used to identify it, would likely benefit greatly from international cooperation.

Implications for research and the dimensional approach

Though research aimed at establishing validity and reliability of specific assessment tools is very important, there may also be a need for further research on how to conceptualise and define mental disorders. In this regard, the DIP is well suited for use in research in areas such as the epidemiological characteristics and genetics of severe mental disorder, as it is directed specifically towards this group of disorders. The DIP also has an advantage in being independent of the current diagnostic systems. This makes it a versatile tool that allows the researcher to obtain both syndrome-oriented estimates based on different theory-driven clusters of symptoms or current diagnostic criteria, and symptom-oriented estimates based on the presence of single symptoms and factors associated with severe mental disorders. This is important when considering alternate ways to define severe mental disorder, as it allows for investigations that are relatively independent from the established definitions.

In light of the lack of a complete consensus about definitions and diagnostic criteria for specific disorders, the advantage that the DIP has by not being bound to any specific diagnostic systems becomes even more apparent. The diagnostic module of the DIP assesses the well-known, individual symptoms and associated illness factors specified in the Operational Criteria for Psychoses (OPCRIT) checklist, and it seems to be easier to obtain agreement between professionals on the presence of individual symptoms and signs, than on complete syndromes. Because of its structure and lack of inherent assumptions about the relationship between individual symptoms and illness factors, the DIP can easily be modified for use in conjunction with any new revisions of diagnostic systems, changes in their definitions of specific diagnoses, or inclusion of a more dimensional approach to mental disorders.

The DIP is suited for a more dimensional approach to mental disorder, in terms of its independent assessment of symptoms that can broadly be characterised as for instance manic, depressive, psychotic, delusional or as hallucinations. This contrasts with an approach that focuses on specific diagnostic groups or narrow diagnostic categories, in an attempt to define separate syndromes. Using the current definition of bipolar disorder as an example, it may

prove more useful to describe the affected individuals in terms of their experience of manic, depressive or psychotic symptoms separately, rather than by diagnoses defined by specific configurations of these symptoms and their relative dominance in the clinical picture. While it is still uncertain to what degree the upcoming revisions of the major diagnostic systems, ICD-11 and DSM-V, will incorporate a more dimensional approach to mental disorders, such an approach would nevertheless be in line with an emerging direction in genetic research, indicating that what is inherited may not be specific syndromes, such as schizophrenia or bipolar disorder, but rather the susceptibility to experience individual phenomena such as psychosis or mania (e.g. Edvardsen et al., 2008; Steinberg et al., 2010). Furthermore, it would be in line with current criteria for use of psychopharmaceutical treatment, where specific medications are targeted at types of symptoms, rather than at a given disorder or diagnosis as a whole. In this regard, the DIP would be useful in obtaining a direct measure of the specific symptoms, in a context not bound to any diagnostic system. Thus, the DIP may prove to be a useful tool in research, both epidemiological and conceptual, because its structure makes it highly versatile in terms of level of measurement and potential independence from current diagnostic systems and the assumptions they are based on.

Implications for clinical practice

There are numerous commonly used Norwegian translations of diagnostic interviews, but to our knowledge, few published reliability studies. This is alarming, as even though a certain diagnostic assessment tool has been shown to be reliable in its original form, the translation procedure may influence it in ways that threaten reliability and validity. Reliability studies are important both in terms of securing the quality of the interview after the translation procedure, and to test the interview in the specific environment in which it is to be used, in this case; the Norwegian population of psychiatric patients. As previously described, DIP is a semi-structured interview that leaves room for the interviewer to make clinical interpretations, while at the same time staying within the frames of the standardized structure. The advantages of semi-structured interviews over fully structured interviews, has already been discussed earlier in this paper.

Given that there are few diagnostic interviews directed specifically towards low prevalent severe mental disorders, and that the use of structured diagnostic approaches has been shown to improve reliability and validity of psychiatric diagnoses, the results of this

study are promising in that the DIP may prove useful in this regard. The results indicate that the Norwegian translation of the DIP may be applicable as a reliable diagnostic interview in the Norwegian population of patients with severe disorders, should the final assessment yield satisfactory validity and reliability. Though there are other interviews that include symptoms of severe mental disorder, the DIP has features that directs it specifically towards thorough assessment of low prevalent disorders, and it can be expected to prove especially useful in making differential diagnoses in the realm of severe affective and psychotic disorders. However, for the purpose of this study, the most important finding is that the DIP has shown good reliability in differentiating bipolar and manic disorders from other affective and psychotic disorders. This is especially important knowing the difficulties experienced by clinicians when trying to differentiate bipolar disorder from disorders with overlapping symptoms, and the relatively higher risk of misdiagnosis of this particular disorder.

Experiences from the study also indicate that the DIP is a user-friendly interview, both for the respondents and the interviewers. All seven raters found the DIP easier to administer than other similar assessment tools such as SCID I, and respondents also reported the DIP to be an understandable assessment tool. This is an important factor in making it more likely that clinicians will apply this type of structured diagnostic tool, as they may be concerned both about how easy the interview is to perform, and to what degree it affects the establishment of a relationship with the patients in this phase of treatment. Furthermore, the DIP may be useful in clinical settings because it allows the clinician to get a direct measure of each symptom, and quickly obtain an overall picture of the patients' symptomatology. Its usefulness can also be seen when measuring symptomatology over time, as the problem with diagnosing bipolar disorder often lies with the clinicians failure to inquire about earlier illness episodes with both depressive and manic symptoms. The direct assessment demanded by the use of structured interviews can contribute to improve the assessment of such episodes. On the other hand, the use of the DIP, which includes use of the accompanying computer software, has the potential to create distance between the clinician and the complete clinical picture, since the diagnoses are generated according to algorithms not obvious to the clinician. For this reason, the use of clinical experience in the final diagnostic evaluation is very important. In addition, close attention must be paid to decisive items when administering the interview. One example of such an item is the relationship between psychotic and affective symptoms, which is given much weight when both types of symptoms are present.

In other interviews, such as SCID I, this type of decisive items may be more apparent during the administration, because the interviewer may then be prompted to move in a different direction when continuing the interview. Thus, while the DIP is easier to administer due to its straightforward structure and order of modules, this shows the importance of adequate experience and training in order to administer the interview in a satisfactory manner.

Another advantage of the DIP is its recommendations to use multiple sources of information, such as case notes or family and friends, in addition to the patient's responses. Meant for use in diagnosing patients with severe mental disorders, the information obtained during the interview may not always be credible, or the respondent may have difficulties in reporting important information, especially if they are in acute illness phases. Thus, the interview can be scored on the basis of several sources, yielding a diagnostic result based on a summary of all available information. The use of multiple sources of information, such as family members, is known to facilitate both the diagnostic process and the validity and reliability of the diagnosis it yields.

Summary and conclusion

In general, this study indicates that the DIP is a reliable tool for the assessment of severe mental disorders, both in the context of current diagnostic classification criteria, and on the level of individual symptoms and associated illness factors. The interview may prove to be a useful tool to reduce misdiagnosis and consequently the risk for delayed, ineffective, or harmful treatment of patients with bipolar disorder (and other severe mental disorders), as well as reduce the time spent before a correct diagnosis is given. The main reasons for this being that it is directed specifically towards low prevalent, severe mental disorders to a greater extent than existing interviews, and that it provides a structured assessment that has been reported to be user friendly by both interviewers and respondents. For these reasons, the DIP is also well suited for research purposes, with its structure allowing for assessment on both diagnostic and symptomatic levels, as well as including associated illness factors. Furthermore, it allows for assessment in a more dimensional context than existing interviews, which are for the most part tied to specific diagnostic systems, and syndrome-oriented rather than symptom-oriented. The fact that the DIP is independent of current diagnostic systems, and more symptom-oriented, is one of its greatest advantages. This makes it a versatile tool that is compatible with, or can easily be updated in accordance with, changes in

conceptualisations of mental disorders that may result from revisions of diagnostic systems and progress in research.

Before it can be concluded that the Norwegian translation of DIP can fill this role in the assessment of severe mental disorders in research and clinical domains, its validity and reliability must be established with a larger sample than what the results of this study are based on. As mentioned, this is currently being done in the larger research project that this study is a part of.

References

- Adler, C. M. & Strakowski, S. M. (2003). Boundaries of schizophrenia. *Psychiatric Clinics of North America*, 26, 1-23.
- Akiskal, H. S. (2008). The emerge of the bipolar spectrum: validation along clinical-epidemiological and familial-genetic lines. *Psychopharmacological Bulletin*, 40, 99-115.
- Altshuler, L. L., Post, R. M., Leverich, G. S., Mikalauskas, K., Rosoff, A., & Ackerman, L. (1995). Antidepressant-induced mania and cycle acceleration: a controversy revisited. *The American Journal of Psychiatry*, 152, 1130-1138.
- Andrews, G., Peters, L., Guzman, A. M., & Bird, K. (1995). A comparison of two structured diagnostic interviews: CIDI and SCAN. *Australian and New Zealand Journal of Psychiatry*, 29, 124-132.
- Angst, J. (2007). The bipolar spectrum. *British Journal of Psychiatry*, 190, 189-191.
- American Psychiatric Association (1980). *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.). American Psychiatric Association: Washington, DC.
- American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders: DSM-III-R* (Rev. 3rd ed.). American Psychiatric Association: Washington, DC.
- American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV* (4th ed.). American Psychiatric Association: Washington, DC.
- Awad, A. G., Rajagopalan, K., Bolge, S. C., & McDonnell, D. D. (2007). Quality of life among bipolar patients misdiagnosed with major depressive disorder. *The Primary Care Companion to the Journal of Clinical Psychiatry*, 9, 195-202.
- Benazzi, F. (2008). A relationship between bipolar II disorder and borderline personality disorder? *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, 32, 1022-1029.
- Berk, M., Hallam, K., Lucas, N., Hasty, M., McNeil, C.A., Conus, P., Kader, L., & McGorry, P. D. (2007). Early intervention in bipolar disorders: opportunities and pitfalls. *The Medical Journal of Australia*, 187, 11-14.
- Birchwood, M., Iqbal, Z., Chadwick, P., & Trower, P. (2000). Cognitive approach to depression and suicidal thinking in psychosis. 1. ontogeny of post-psychotic depression. *British Journal of Psychiatry*, 177, 516-521.

- Brugha, T. S., Jenkins, R., Taub, N., Meltzer, H., & Bebbington, P. E. (2001). A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). *Psychological Medicine, 31*, 1001-1013.
- Buckley, P. F., Gowans, A., Sebastian, C. S., Pathiraja, A., Brimeyer, A., & Stirewalt, E. (2004). The boundaries of schizophrenia: overlap with bipolar disorder. *Current Psychosis & Therapeutics Reports, 2*, 49-56.
- Castle, D. J. & Jablensky, A. (2005). Diagnosis and classification in psychiatry. In P. Wright, J. Stern, & M. Phelan (eds) *Core Psychiatry*. Elsevier Saunders: Edinburgh, pp. 507-515.
- Castle, D., Jablensky, A., McGrath, J., Carr, V., Morgan, V., Watterreus, A., Valuri, G., Stein, H., McGruffin, P., & Farmer, A. (2006). The diagnostic interview for psychoses (DIP): development, reliability and applications. *Psychological Medicine, 36*, 69-80.
- Cheng, A. T. A., Tien, A. Y., Chang, C. J., Brugha, T. S., Cooper, J. E., Lee, C. S., Compton, W., Liu, C. Y., Yu, W. Y., & Chen, H. M. (2001). Cross-cultural implementation of a Chinese version of the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) in Taiwan. *British Journal of Psychiatry, 178*, 567-572.
- Cho, M. J., Hahm, B. J., Suh, D. W., Hong, J. P., Bae, J. N., Kim, J. K., Lee, D. W., & Cho, S. J. (2002). Development of a Korean version of the Composite International Diagnostic Interview (K-CIDI). *Journal of Korean neuropsychiatric association, 41*, 123-137.
- Cicchetti, D. V. & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*, 551-558.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.
- Craddock, N., Jones, I., Kirov, G., & Jones, L. (2004). The bipolar affective disorder dimension scale (BADDS) - a dimensional scale for rating lifetime psychopathology in bipolar spectrum disorders. *BMC psychiatry, 4*.
- Edvardsen, J., Torgersen, S., Røysamb, E., Lygren, S., Skre, I., Onstad, S & Øien, P. (2008). Heritability of bipolar spectrum disorders. Unity or heterogeneity? *Journal of Affective Disorders, 106*, 229-240.

- Farmer, A., S. Wessley, S., Castle, D., & McGuffin, P. (1992). Methodological issues in Using a Polydiagnostic Approach to Define Psychotic Illness. *British Journal of Psychiatry*, 161, 824-830.
- Feinstein, A. R. & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43,543-549.
- Fleiss, J. L. & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33, 613-619.
- Ghaemi, S. N., Boiman, E. E., & Goodwin, F. K. (2000). Diagnosing bipolar disorder and the effect of antidepressants: A naturalistic study. *Journal of Clinical Psychiatry*, 61, 804-808.
- Ghaemi, S. N., Sachs, G. S., Chiou, A. M., Pandurangi, A. K., & Goodwin, F. K. (1999). Is bipolar disorder still underdiagnosed? Are antidepressants overutilized? *Journal of Affective Disorders*, 52, 135-144.
- Ghaemi, S. N., Stoll, A. L., & Pope, H. G. (1995). Lack of insight in bipolar disorder. *Journal of Nervous and Mental Disease*, 183, 464-467.
- Glick, I. D. (2004). Undiagnosed bipolar disorder: new syndromes and new treatments. *The Primary Care Companion to the Journal of Clinical Psychiatry*, 6, 27-31.
- Goodwin, F. K. & Jamison, K. R. (2007) *Manic-depressive illness. Bipolar disorders and recurrent depression*. Oxford university press, New York.
- Goldberg, J. F. & Ernst, C. L. (2002). Features associated with the delayed initiation of mood stabilizers at illness onset in bipolar disorder. *Journal of Clinical Psychiatry*, 63, 985-991.
- Hesselbrock, M., Easton, C., Bucholz, K. K., Schuckit, M., & Hesselbrock, V. (1999). A validity study of the SSAGA – a comparison with the SCAN. *Addiction*, 94, 1361-1370.
- Hirschfeld, R. M. A. (2001). Bipolar spectrum disorder: improving its recognition and diagnosis. *Journal of Clinical Psychiatry*, 62, 5-9.
- Hirschfeld, R. M. A. (2002). The mood disorder questionnaire: A simple, patient-rated instrument for bipolar disorder. *The Primary Care Companion to the Journal of Clinical Psychiatry*, 4, 9-11.

- Hirschfeld, R. M. A., Lewis, L., & Vornik, L. A. (2003). Perceptions and impact of bipolar disorder: how far have we really come? Results of the National Depressive and Manic-Depressive Association 2000 survey of individuals with bipolar disorder. *Journal of Clinical Psychiatry, 64*, 161-174.
- Jablensky, A., McGrath, J., Herrman, H., Castle, D., Gureje, O., Evans, M., Carr, V., Morgan, V., Korten, A., & Harvey, C. (2000). Psychotic disorders in urban areas: an overview of the Study on Low Prevalence Disorders. *Australian and New Zealand Journal of Psychiatry, 34*, 221-236.
- Jensen, P., Roper, M., Fisher, P., Piacentini, J., Canino, G., Richters, J., Rubio-Stipec, M., Dulcan, M., Goodman, S., Davies, M., Rae, D., Shaffer, D., Bird, H., Lahey, B., & Schwab-Stone, M. (1995). Test-retest reliability of the Diagnostic Interview Schedule for Children (DISC 2.1). *Archives of General Psychiatry, 52*, 61-71.
- Kadri, N., Agoub, M., El Gnaoui, S., Alami, K. M., Hergueta, T., & Moussaoui, D. (2005). Moroccan colloquial Arabic version of the Mini International Neuropsychiatric Interview (MINI): qualitative and quantitative validation. *European Psychiatry, 20*, 193-195.
- Kendell, R. & Jablensky, A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *The American Journal of Psychiatry, 160*, 4-12.
- Kessler, R. C., Akiskal, H. S., Angst, J., Guyer, M., Hirschfeld, R. M. A., Merikangas, K. R., & Stang, P. E. (2006). Validity of the assessment of bipolar spectrum disorders in the WHO CIDI 3.0. *Journal of Affective Disorders, 96*, 259-269.
- Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H. U., & Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the National Comorbidity Survey. *Archives of General Psychiatry, 51*, 8-19.
- Kringlen, E. (2002). Trekk fra internasjonal psykiatrisk epidemiologi. *Norsk Epidemiologi, 12*, 173-179.
- Landis, J. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.
- Laursen, T. M., Agerbo, E., & Pedersen, C. B. (2009). Bipolar disorder, schizoaffective disorder, and schizophrenia overlap: a new comorbidity index. *The Journal of Clinical Psychiatry, 70*, 1432-1438.

- Lecrubier, Y., Sheehan, D. V., Weiller, E., Amorim, P., Bonora, I., Sheehan, K. H., Janavs, J., & Dunbar, G.C. (1997). The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. *European Psychiatry, 12*, 224-231.
- Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., & Hultman, C. M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet, 373*, 234-239.
- Lish, J. D., Dime-Meenan, S., Whybrow, P. C., Price, R. A., & Hirschfeld, R. M. A. (1994). The National Depressive and Manic-depressive Association (DMDA) survey of bipolar members. *Journal of Affective Disorders, 31*, 281-294.
- Lora, A., Bezzi, R., & Erlicher, A. (2007). Estimating the prevalence of severe mental illness in mental health services in Lombardy (Italy). *Community Mental Health Journal, 43*, 341-357.
- McGuffin, P., Farmer, A., & Harvey, I. (1991). A polydiagnostic application of operational criteria in studies of psychotic illness. Development and reliability of the OPCRIT system. *Archives of General Psychiatry, 48*, 764-770.
- MacKinnon, A. (2000). A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement. *Computers in Biology and Medicine, 30*, 127-134.
- MacKinnon, D. F. & Pies, R. (2006). Affective instability as rapid cycling: theoretical and clinical implications for borderline personality and bipolar spectrum disorder. *Bipolar Disorders, 8*, 1-14.
- Marneros, A. (2003). Schizoaffective disorder: clinical aspects, differential diagnosis, and treatment. *Current Psychiatry Reports, 5*, 202-205.
- Marneros, A. & Angst, J. (2000). *Bipolar disorders 100 years after manic depressive insanity*. Kluwer Academic Publishers: Dordrecht.
- Matza, L. S., Rajagopalan, K. S., Thompson, C. L., & Lissovoy, G. (2005). Misdiagnosed patients with bipolar disorder: comorbidities, treatment patterns, and direct treatment costs. *The Journal of Clinical Psychiatry, 66*, 1432-1440.
- McElroy, S. L., Altshuler, L. L., Suppes, T., Keck, P. E. Jr., Frye, M. A., Denicoff, K. D., Nolen, W. A., Kupka, R. W., Leverich, G. S., Rochussen, J. R., Rush, A. J., & Post, R. M. (2001). Axis I psychiatric comorbidity and its relationship to historical illness

- variables in 288 patients with bipolar disorder. *The American Journal of Psychiatry*, 158, 420-426.
- Miller, P. R., Dasher, R., Collins, R., Griffiths, P., & Brown, F. (2001). Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews. *Psychiatry Research*, 105, 255-264.
- Mork, E., Mehlum, L., & Walby, F. A. (2009). Selvmord ved depresjon med psykotiske symptomer og bipolar lidelse: forekomst, risikofaktorer og neurobiologiske forhold. *Suicidalogi*, 14, 8-12.
- NOU, Helse- og omsorgsdepartementet (1999:13). *Kvinnens helse i Norge*. Retrieved 23.04.10 from <http://www.regjeringen.no/nb/dep/hod/dok/nouer/1999/nou-1999-13.html?id=141704>
- Piver, A., Yatham, L. N., & Lam, R. W. (2002). Bipolar spectrum disorders. New perspectives. *Canadian Family Physician*, 48, 896-904.
- Pope, H. G. & Lipinski, J. F. (1978). Diagnosis in schizophrenia and manic-depressive illness: a reassessment of the specificity of schizophrenic symptoms in the light of current research. *Archives of General Psychiatry*, 35, 811-828.
- Quintana, M. I., Gastal, F. L., Jorge, M. R., Miranda, C. T., & Andreoli, S. B. (2007). Validity and limitations of the Brazilian version of the Composite International Diagnostic Interview (CIDI 2.1). *Revista Brasileira de Psiquiatria*, 29, 18-22.
- Regier, D. A., Kaelber, C. T., Rae, D. S., Farmer, M. E., Knauper, B., Kessler, R. C., & Norquist, G. S. (1988). Limitations of diagnostic criteria and assessment instrument for mental disorders. *Archives of General Psychiatry*, 55, 109-115.
- Rijnders, C. A. T., van den Berg, J. F. M., Hodiament, P. P. G., Nienhuis, F. J., Furer, J. W., Mulder, J., & Giel, R. (2000). Psychometric properties of the schedules for clinical assessment in neuropsychiatry (SCAN-2.1). *Social Psychiatry and Psychiatric Epidemiology*, 35, 348-352.
- Robins, L. N. (1985). Epidemiology: reflections on testing the validity of psychiatric interviews. *Archives of General Psychiatry*, 42, 918-924.
- Rogers, R. (2001). *Handbook of diagnostic and structured interviewing*. The Guilford Press: New York.
- Rogers, R. (2003). Standardizing DSM-IV diagnoses: the clinical applications of structured interviews. *Journal of Personality Assessment*, 81, 220-225.

- Rossi, A., Morgan, V., Amaddeo, F., Sandri, M., Grigoletti, L., Maggioni, F., Ferro, A., Rigon, E., Donisi, V., Venturi, V.V., Goria, F., Skre, I., Tansella, M., & Jablensky, A. (2010). Diagnosing psychotic disorders: validity, reliability and applications of the Diagnostic Interview for Psychosis (DIP). Italian version. *Epidemiologia e Psichiatria Sociale*, *19*, 33-43.
- Ruggeri, M., Leese, M., Thornicroft, G., Bisoffi, G., & Tansella, M. (2000). Definition and prevalence of severe and persistent mental illness. *British Journal of Psychiatry*, *177*, 149-155.
- Sachs, G. S. & Thase, M. E. (2000). *Bipolar disorder: a systematic approach to treatment*. Martin Dunitz Ltd: London.
- Sartorius, N., Kaelber, C. T., Cooper, J. E., Roper, M. T., Rae, D. S., Gulbinat, W., Üstün, T. B., & Regier, D. A. (1993). Progress toward achieving a common language in psychiatry. Results from the field trial of the clinical guidelines accompanying the WHO classification of mental and behavioral disorders in ICD-10. *Archives of General Psychiatry*, *50*, 115-124.
- Schulze, T. G, Hedeker, D., Zandi, P., Rietschel, M., & McMahon, F. J. (2006). What is familial about familial bipolar disorder? *Archives of General Psychiatry*, *63*, 1368-1376.
- Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of clinical psychiatry*, *59*, 22-33.
- Shi, L., Thiebaud, P., & McCombs, J. S. (2004). The impact of unrecognized bipolar disorders for patients treated for depression with antidepressants in the fee-for-services Californial Medicaid (Medi-Cal) program. *Journal of Affective Disorders*, *82*, 373-383.
- Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428.
- Skre, I., Onstad, S., Torgersen, S., & Kringlen, E. (1991). High interrater reliability for the Structured Clinical Interview for DSM -III-R Axis I (SCID – I). *Acta Psychiatrica Scandinavia*, *84*, 167-173.

- Spitzer, R. L., Williams, J. B., Gibbon, M., & First, M. B. (1992). The structured clinical interview for DSM-III-R (SCID). I: history, rationale, and description. *Archives of General Psychiatry*, *49*, 624-629.
- Steinberg, S., Mors, O., Børglum, A. D., Gustafsson, O., Werge, T., Mortensen, P. B., ... Stefansson, K. (2010). Expanding the range of ZNF804A variants conferring risk of psychosis. *Molecular Psychiatry*. Advance online publication. doi: 10.1038/mp.2009.149
- Suppes, T., Leverich, G. S., Keck Jr., P. E., Noen, W. A., Denicoff, K. D., Altshuler, L. L., McElroy, S. L., Rush, A. J., Kupka, R., Frye, M. A., Bickel, M., & Post, R. M. (2001). The Stanley Foundation Bipolar Treatment Outcome Network II. Demographics and illness characteristics of the first 261 patients. *Journal of Affective Disorders*, *67*, 45-59.
- Swann, A. C., Bowden, C. L., Calabrese, J. R., Dilsaver, S. C., & Morris, D. D. (1999). Differential effect of number of previous episodes of affective disorder on response to lithium or divalproex in acute mania. *The American Journal of Psychiatry*, *156*, 1264-1266.
- Swann, A. C., Steinberg, J. L., Lijffijt, M., & Moeller, F. G. (2009). Continuum of depressive and manic mixed states in patients with bipolar disorder: quantitative measurement and clinical features. *World Psychiatry*, *8*, 166-172.
- Taiminen, T., Ranta, K., Karlsson, H., Lauerma, H., Leinonen, K. M., Wallenius, E., Kaljonen, A., & Salokangas, R. K. R. (2001). Comparison of clinical and best-estimate research DSM-IV diagnoses in a Finnish sample of first-admission psychosis and severe affective disorder. *Nordic Journal of Psychiatry*, *55*, 107-111.
- Taylor, M. A., Gaztanaga, P., & Abrams, R. (1974). Manic-depressive illness and acute schizophrenia: a clinical, family history, and treatment-response study. *The American Journal of Psychiatry*, *131*, 678-682.
- Thompson W.D. & Walter S.D. (1988) A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology*, *41*, 949-58.
- Tondo, L., Isacson, G. & Baldessarini, R.J. (2003). Suicidal behavior in bipolar disorder: Risk and prevention. *CNS drugs*, *17*, 7, 491-511.
- Torgersen, S., Cramer, V., & Kringlen, E. (2002). Psykiske forstyrrelser i ulike bydelsregioner i Oslo. *Norsk Epidemiologi*, *12*, 265-268.

- Torrey, E. F. (1999). Epidemiological comparison of schizophrenia and bipolar disorder. *Schizophrenia Research*, 39, 101-106.
- University of Western Australia, School of Psychiatry & Clinical Neurosciences, Centre for Clinical Research in Neuropsychiatry (2008). Diagnostic Interview for Psychoses. Diagnostic Module (DIP-DM). Version Nov. 2008, DIP 4.0.11. Contact: john.dean@health.wa.gov.au.
- Van Snellenberg, J. X. & de Candia, T. (2009). Meta-analytic evidence for familial coaggregation of schizophrenia and bipolar disorder. *Archives of General Psychiatry*, 66, 148-155.
- World Health Organization (1993). *The ICD-10 classification of mental and behavioural disorders. Diagnostic criteria for research*. World Health Organization: Geneva.
- Widiger, T. A. & Samuel, D. B. (2005). Diagnostic categories or dimension? A question for the Diagnostic and Statistical Manual of Mental Disorders – Fifth Edition. *Journal of Abnormal Psychology*, 114, 494-504.
- Williams, J., Farmer, A. E., Ackenheil, M., Kaufmann, C. A., McGuffin, P., & the OPCRIT Reliability Research Group (1996). A multicentre inter-rater reliability study using the OPCRIT computerized diagnostic system. *Psychological Medicine*, 26, 775-783.
- Øiesvold, T., Nivison, M., Østensen, L., Hansen, V., Sørgaard, K. W., & Skre, I. (submitted). The validity of administrative registers in psychiatric hospital with special emphasis on bipolar disorder.
- Ösby, U., Brandt, L., Correia, N., Ekborn, A., & Sparén, P. (2001). Excess mortality in bipolar and unipolar disorder in Sweden. *Archives of General Psychiatry*, 58, 844-850.

APPENDIX A

The Diagnostic Module of the Diagnostic Interview for Psychoses (DIP-DM)

DIP-DM consists of 97 items in the following areas and order:

- General items
- Pre-morbid characteristics and onset
- Family history
- Depression
- Mania
- Hallucinations
- Subjective thought disorder
- Delusions
- Insight
- Response to medication
- General rating on psychotic symptoms
- Substance use: alcohol; non-medical use of drugs
- Alcohol and drug abuse and dependence
- Duration and course
- Behaviour
- Affect
- Speech

APPENDIX B

Diagnostic Classification Systems Compatible with the DIP-DM

- ICD-10
- DSM-IV
- DSM-III-R
- Research Diagnostic Criteria
- DSM-III
- Feighner et al.
- Carpenter FRS
- French classification
- Taylor Abrams
- Tsuang-Winokur subtyper
- Crow subtyper
- Farmer subtyper

APPENDIX C

ICD-10 and DSM-IV Sub-classification of Bipolar Disorder

ICD-10 classification of bipolar disorder

- F 31.0 Current episode hypomanic
- F 31.1 Current episode manic without psychotic symptoms
- F 31.2 Current episode manic with psychotic symptoms
 - 31.20 With mood-congruent psychotic symptoms
 - 31.21 With mood-incongruent psychotic symptoms
- F 31.3 Current episode mild or moderate depression
 - 31.30 Without somatic symptoms
 - 31.31 With somatic symptoms
- F 31.4 Current episode severe depression without psychotic symptoms
- F 31.5 Current episode severe depression with psychotic symptoms
- F 31.6 Current episode mixed
- F 31.7 Currently in remission
- F 31.8 Other bipolar affective disorders

DSM-IV classification of bipolar disorder

- 296.0x Bipolar I disorder, single manic episode
- 296.40 Bipolar I disorder, most recent episode hypomanic
- 296.4x Bipolar I disorder, most recent episode manic
- 296.6x Bipolar I disorder, most recent episode mixed
- 296.5x Bipolar I disorder, most recent episode depressed
- 296.7 Bipolar I disorder, most recent episode unspecified
- 296.89 Bipolar II disorder
- 296.80 Bipolar disorder not otherwise specified
- 301.13 Cyclothymic disorder

APPENDIX D

Sample from the Norwegian Translation of the DIP

MANI**40. Hevet stemningsleie (OPCRIT 35)**

Ekspansivt (hevet) stemningsleie (SCAN 10.001)

- **Jeg har nå stilt deg noen spørsmål om depresjon; nå vil jeg spørre deg om du har følt det motsatte av depresjon, for eksempel vært intenst glad eller oppløftet, uten noen spesiell grunn?**

Om det foreligger bevis for en nåværende stemningslidelse, still spørsmålene under ordrett; om det foreligger bevis for tidligere episode, tilpass spørsmålene deretter. Om det er flere enn en episode, intervju på grunnlag av den mest alvorlige episoden.

- Så oppløftet at det var unaturlig?
- Kan du beskrive den følelsen?
- Var den lite typisk for deg?
- Hvor lenge varte det? Dager? Mer enn en uke?
- Hadde du ruset deg for å bli "høy" ?

0 = Ikke tilstede

1 = Tilstede i minst fire dager

2 = Tilstede i minst en uke

3 = Tilstede i minst to uker ELLER om det vedvarte < en uke, respondenten ble innlagt på sykehus for en affektiv lidelse

Respondentens dominerende stemningsleie er hevet og ute av proporsjon til omstendighetene. Respondenten er euforisk eller i et hevet stemningsleie mesteparten av tiden. Det er viktig å finne ut av om dette er tilbakevendelse til en "normal" tilstand for en person som har vært deprimert.

Det skal ikke skåres:

Forbigående " godt humør".

Rus-indusert hevet stemningsleie eller ekspansivitet.

40. Hevet stemningsleie

PS

PY

LT

41. Irritabelt humør (OPCRIT 36)**Irritabelt humør (SCAN 10.002)**

- **Nå vil jeg spørre deg om du har følt deg veldig irritabel eller overdrevent irritert overfor andre, slik at du ofte har mistet besinnelsen?**
- Har andre personer kommentert eller sagt at du er altfor utålmodig?
- Hvor lenge har du følt det slik?

0 = Ikke tilstede

1 = Tilstede i minst fire dager

2 = Tilstede i minst en uke

3 = Tilstede i minst to uker ELLER hvis det vedvarte < en uke ble respondenten innlagt på sykehus for en affektiv lidelse

*Respondentens humør er hovedsakelig irritabelt.**Respondenten beskriver humøret sitt som hissig og oppfarende**Dette kan beskrives som:*

- *Vedvarende sinne; eller*
- *Utålmodighet; eller*
- *Å være "på hugget" og klar til å hisse seg opp over små irritasjoner; eller*
- *Har kort lunte,*

*Som varer i minst en uke. Personen kan erkjenne at reaksjonene er overdrevne, ute av proposjon i forhold til omstendighetene og vanskelige å kontrollere. Det kan også være en ubehagelig opplevelse.**Det skal ikke skåres:**Ved enkeltstående episode med irritasjon eller tap av beherskelse.***41. Irritabelt humør**

--	--	--

PS

PY

LT

Om svaret er NEI på både hevet stemningsleie (punkt 40) og irritert humør (punkt 40): gå direkte til punkt 49: hallusinasjoner

Skår punkt 42 til 48 i forhold til episodene med hevet stemningsleie eller mani, som kommer frem gjennom punkt 40 og 41.

42. Tankeflukt (OPCRIT 31)

Påtrengende og flyktige tanker (SCAN 10.004)

- **Føler du at tankene dine myldrer og raser gjennom hodet ditt?**
- Slik at du ikke kan holde følge med dem?
- Kan du beskrive det?
- Hvor lenge har det holdt på?

- 0 = Ikke tilstede
- 1 = Tilstede i minst fire dager
- 2 = Tilstede i minst en uke
- 3 = Tilstede i minst to uker

Respondenten opplever tanker som raser gjennom hans/hennes hode. Bilder og ideer flyr raskt gjennom hodet, den ene tanken tar den andre. Andre kan observere ideflukt og ha problemer med å følge hva respondenten sier på grunn av hurtig tale og taleflom.

42. Tankeflukt

--	--	--

PS PY LT

43. Distraherbarhet (OPCRIT 21)

Distraherbarhet (SCAN 10.006)

- **Har du lett blitt avledet av uvesentlige ting som skjer rundt deg?**
- Har du klart å holde oppmerksomheten din på en ting lenge nok til å håndtere det skikkelig?
- Hvor lenge har det vært slik?

- 0 = Ikke tilstede
- 1 = Tilstede i minst fire dager
- 2 = Tilstede i minst en uke
- 3 = Tilstede i minst to uker

Respondenten opplever problemer med å konsentrere seg om hva som skjer rundt han/henne. Deres tankeflyt og tale forstyrres ofte fordi deres oppmerksomhet lett retter seg mot urelevante og uvesentlige faktorer i nære omgivelser.

43. Distraherbarhet

--	--	--

PS PY LT

APPENDIX E

Written Consent Form and Accompanying Information

Forespørsel om deltakelse i forskningsprosjektet***”En norsk validering av The diagnostic interview for psychosis (DIP)”***

Du har nettopp blitt intervjuet med en norsk utgave av et engelsk intervju, DIP. De opplysningene som kom fram i intervjuet vil bli skrevet inn i din journal, og vil bidra til at avdelingen stiller en mest mulig riktig diagnose, som igjen gir grunnlag for å gi deg bedre behandling. Intervjuet kommer derfor direkte til nytte for deg.

Bakgrunn og hensikt

Vi vil nå spørre deg om du er villig til å delta i et forskningsprosjekt vi har om dette intervjuet. Man gjør slike intervjuer for å få en grundig kartlegging av symptomer og plager pasienter har. Intervjuet utreder symptomer på psykose, oppstemthet, nedstemthet og rusproblemer, samt en del generelle spørsmål. Hensikten med prosjektet er å komme fram til en norsk oversettelse som er vitenskapelig like god som den engelske originalen.

Hva innebærer studien?

Du kan samtykke til 2 forskjellige nivåer av deltagelse:

- 1) Samtykke til at de opplysningene som allerede er samlet inn under intervjuet, blir lagt inn i en separat forskningsdatabase, i aidentifisert form.
- 2) Samtykke til å bli intervjuet 1 gang til, noe senere, av en annen lege, psykolog eller psykologistudent, enn den som gjorde det første intervjuet. Dette intervjuet kan bli det samme som det første (DIP), eller et annet som kalles SCID.

Når vi gjør 2 intervju med DIP, er det for å sjekke at 2 forskjellige intervjuere kommer til samme resultat. Når vi gjør ett intervju med DIP og senere ett med SCID, er det for å sjekke om DIP er like bra som SCID, som er et intervju som har vært lenge i bruk i psykiatrien.

Mulige fordeler og ulemper ved deltagelse

Du vil ikke få direkte nytte av å delta i selve forskningsprosjektet. Men du vil indirekte få nytte av at vi får laget et vitenskapelig godt norsk intervju, som du kan bli intervjuet med dersom du senere skulle ha behov for hjelp for psykiske plager i helsevesenet. Du vil også ha bidradd til at andre pasienter kan bli intervjuet med et godt intervju i fremtiden.

Du vil få velge mellom 10 Flaxlodd eller kr. 250,- kontant som kompensasjon for bruk av din tid for å delta i studien

Ulempen ved deltagelse er at du vil bruke noen timer på intervjuene, og at opplysninger du allerede har gitt vil måtte gjentas.

Hva skjer med informasjonen om deg?

Informasjonen som registreres om deg skal kun brukes slik som beskrevet i hensikten med studien. Alle opplysningene vil bli behandlet uten navn og fødselsnummer eller andre direkte gjenkjennerende opplysninger. En kode knytter deg til dine opplysninger og gjennom en navneliste, slik av vi kan knytte opplysningene fra de 2 intervjuene sammen.

Navnelisten som kobler navn og kodenummer sammen vil bli oppbevart i låst stålskap hos den overlegen ved institusjonen som er medarbeider i prosjektet (Vidje Hansen ved UNN og Terje Øiesvold ved Nordlandssykehuset). Det er kun prosjektleder som skal ha adgang til navnelisten og som kan finne tilbake til deg. Det vil ikke være mulig å identifisere deg i resultatene av studien når disse blir publisert i et vitenskapelig tidsskrift .

Frivillig deltagelse

Det er frivillig å delta i studien. Du får 24 timers betenkningstid med å svare. Du kan når som helst og uten å oppgi noen grunn trekke ditt samtykke til å delta i studien. Dette vil ikke få konsekvenser for din videre behandling. Dersom du ønsker å delta, undertegner du samtykkeerklæringen på siste side. Om du nå sier ja til å delta, kan du senere trekke tilbake ditt samtykke uten at det påvirker din øvrige behandling. Dersom du senere ønsker å trekke deg eller har spørsmål til studien, kan du kontakte førsteamanuensis Ingunn Skre, Universitetet i Tromsø tlf 77645446, avd.overlege Vidje Hansen, UNN, tlf 77627651, eller overlege Terje Øiesvold, Nordlandssykehuset, tlf 75 50 1524.

Kapittel A- utdypende forklaring av hva studien innebærer

- Kriterier for deltagelse er at man er over 18 år, at man behersker norsk språk, og at man ikke lider av demens eller mental retardasjon.
- Bakgrunnsinformasjon om studien: Korrekt diagnostikk av alvorlige psykiske lidelser som schizofreni og bipolar lidelse er avgjørende og veiledende for valg av behandling og rehabilitering av en sårbar pasientgruppe. Å være i stand til å gjenkjenne og skille mellom positive og negative symptomer på schizofreni, og korrekt diagnostikk av humørlidelser, særlig bipolar lidelse, er problematisk og et område der spesialisthelsetjenesten ofte svikter. Innen psykisk helsevern kan diagnostisk reliabilitet defineres som enigheten mellom ulike klinikere som anvender et felles regelsett for klinisk vurdering. De som foretar denne studien har erfaring med bruk av ulike diagnostiske intervjuer og ønsker å prøve ut om dette nye intervjuet kan dekke et behov for god diagnostikk av alvorlige psykiske lidelser.
- Prosedyre: De pasienter hvor ansvarlig behandler mener det er behov for intervju med DIP for å komme fram til en riktig diagnose, blir intervjuet av 2 personer under oppholdet på avdelingen. Den ene vil gjøre intervjuet, og skåre i DIP, og den andre vil gjøre det samme. Begge disse vil være enten psykolog, lege, psykiater eller psykologstudent. Den som er tilhører gjør en egne vurdering av de opplysningene som fremkommer i intervjuet. De to skåringene sammenlignes så for å undersøke

hvor stor enigheten er om de vurderinger som foretas. Kun de intervjuene der pasienten gir sitt samtykke vil inngå i forskningsmaterialet. Etter at pasienten er blitt forespurt, gis det 24 timer betenkningstid.

- Det er ikke nødvendig for prosjektet at alle som har sagt ja til pkt. 1 i prosedyren, blir intervjuet flere ganger. Vi vil allikevel spørre alle, da vi har erfaring for at mange vil falle fra pga. praktiske hensyn, slik som at de blir utskrevet før vi rekker de senere intervjuene.
- Intervjuet foretas så snart det er praktisk mulig og tilrådelig etter inntak i avdelingen, og intervju nr.2 ønskes foretatt kort tid etter det første. Det nye intervjuet med SCID vil foretas minimum en uke senere enn det første, og ellers så snart som det er mulig eller tilrådelig.

Kapittel B - Personvern, biobank, økonomi og forsikring

Personvern

Opplysninger som registreres om deg er kjønn, alder, og de diagnostiske opplysninger som fremkommer i intervjuet i kodet form. Studien er et samarbeid mellom Universitetet i Tromsø, UNN HF, Nordlandsykehuset, Universitetet i Verona, Italia, og Universitetet i Vestre Australia, Perth, Australia.

Universitetet i Tromsø ved administrerende direktør er databehandlingsansvarlig. Aidentifiserte data vil oppbevares ved Universitetet i Tromsø. Data vil oppbevares i inntil 5 år, eller kortere tid dersom prosjektet avsluttes før. Kun prosjektleder og prosjektmedarbeidere har tilgang til data.

Utlevering av materiale og opplysninger til andre

Hvis du sier ja til å delta i studien, gir du også ditt samtykke til at aidentifiserte opplysninger utleveres til Universitetet i Verona, Italia og til Universitetet i Vestre Australia, Perth, Australia

Rett til innsyn og sletting av opplysninger om deg og sletting av prøver

Hvis du sier ja til å delta i studien, har du rett til å få innsyn i hvilke opplysninger som er registrert om deg. Du har videre rett til å få korrigert eventuelle feil i de opplysningene vi har registrert. Dersom du trekker deg fra studien, kan du kreve å få slettet alle de innsamlede opplysninger, med mindre opplysningene allerede er inngått i analyser eller brukt i vitenskapelige publikasjoner.

Økonomi og (Helse Nord's) rolle

Studien er søkt finansiert gjennom forskningsmidler fra Helse Nord.

Forsikring

Som pasient ved avdelingen omfattes pasienten av institusjonens pasientskadeforsikring.

Informasjon om utfallet av studien

Deltakere har rett til informasjon om resultatene av studien.

Samtykke til deltakelse i studien

1) Jeg er villig til å delta i studien på den måten at data som allerede er innsamlet inkluderes i forskningsdatabasen

(Signert av prosjektdeltaker, dato)

Eller

- Jeg er villig til å delta som nevnt over, pluss bli intervjuet en gang til.
-

APPENDIX F

Items Rated on a Dichotomous (item 14) and an Ordinal (item 46) Scale

<p>14. Poor premorbid work adjustment (OPCRIT 9)</p> <ul style="list-style-type: none"> • Tell me about jobs you had before you first became ill? • What was the longest time you worked in one job before you first became ill? • (If student ask about studies; if housewife, ask about standars of housework) <p>0 = Good premorbid work adjustment 1 = Poor premorbid work adjustment</p> <p><i>Refers to work history before onset of illness. Poor work adjustment = If working and unable to keep any job for more than six months, had a history of frequent changes in job or was only able to sustain a job well below that expected by his/her educational level or training at time of first psychiatric contact. If housewife and persistently very poor standards of housework. If student and badly failing to keep up with studies.</i></p>	<p>14. Premorbid work</p> <p><input type="checkbox"/></p>
<p>46. Reckless activity (OPCRIT 20) Actions based on expansive mood (SCAN 10.012)</p> <ul style="list-style-type: none"> • Have you spent a lot more money than usual during [the PERIOD]? • Have any problems arisen? Do some people think you have been unwise? • Have you done things you later regret? • Have there been troubles in other ways, such as reckless driving? • How long has this been a problem? <p>0 = Not present 1 = Present for at least four days 2 = Present for at least one week 3 = Present for at least two weeks</p> <p><i>The respondent is excessively involved in activities with high potential for painful consequences which is not recognised, e.g. excessive spending and inappropriate gifts to charity, sexual indiscretions, reckless driving, gambling, unaccustomed drunkenness. Such actions may or may not be socially embarrassing.</i></p>	<p>46. Reckless activity</p> <p><input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/></p> <p>PS PY LT</p>