Diagnostic Interview for Genetic Studies (DIGS): Inter-rater and test-retest reliability and validity in a Spanish population

Miguel Roca a,*, Rocio Martin-Santos b, Jerónimo Saiz c, Jordi Obiols d, Maria J. Serrano a, Marta Torrens b, Susana Subirà d, Margalida Gili a, Ricard Navínés b, Angela Ibañez c, Marcos Nadal a, Nieves Barrantes d, Francesca Cañellas e

a Hospital Joan March, University of Balearic Islands, Institut Universitari d’Investigació en Ciències de la Salut, Edifici Guillem Cifre, Carretera de Valldemossa km 7.5, 07122 Palma de Mallorca, Baleares, Spain
b Pharmacologic Research Unit, Institut Municipal d’Investigació Mèdica (IMIM) and Drug Abuse and Psychiatry Department (IAPS), Hospital Universitari del Mar, Barcelona

Received 30 June 2006; received in revised form 13 October 2006; accepted 19 October 2006
Available online 26 December 2006

Abstract

Objective. — To test the reliability and validity of the DIGS in Spanish population.

Methods. — Inter-rater and test-retest reliability of the Spanish version of DIGS was tested in 95 inpatients and outpatients. The resultant diagnoses were compared with diagnoses obtained by the LEAD (Longitudinal Expert All Data) procedure as “gold standard”. The kappa statistic was used to measure concordance between blind inter-raters and between the diagnoses obtained by LEAD procedure and through the DIGS.

Results. — Overall kappa coefficient for inter-rater reliability was 0.956. The kappa value for individual diagnosis varied from major depression = 0.877 to schizophrenia = 1. Test-retest reliability was 0.926. Kappa for all individual target diagnoses ranged from 0.776 (major depression) to 1. Kappa between LEAD procedure and DIGS ranged from 0.704 (major depression) to 0.825 (bipolar I disorder).

Conclusion. — Most of the DSM-IV major psychiatric disorders can be assessed with acceptable to excellent reliability with the Spanish version of the DIGS interview. The Spanish version of DIGS showed an acceptable to excellent concurrent validity. Giving the good reliability and validity of Spanish version of DIGS it should be considered to identify psychiatric phenotypes for genetics studies.

Keywords: DIGS; Genetic studies; Validation

1. Introduction

Genetic studies in psychiatry have increased considerably in recent years. The results obtained are contradictory due to the fact that mental illnesses are complex and do not follow a Mendelian inheritance. These contradictory results can be attributed to the limited specificity of diagnostic phenotypes in psychiatry [2,10] with frequently overlapping symptomatology and unreliable diagnostic instruments [8].

Using standard diagnostic criteria and semi-structured interviews has reduced the risk of inadequate assessment. However, the main diagnostic interviews used in psychiatry were not designed for genetic studies. In 1994 the National Institute of Mental Health (NIMH) developed a specific instrument for genetic studies: the Diagnostic Interview for Genetic Studies (DIGS) [12]. The DIGS has polydiagnostic capacity, enables a detailed assessment of the course of the illness, chronology of the affective and psychotic disorders and comorbidity, as
well as an additional description of symptoms including the possibility of an algorithmic scoring. The DIGS includes a section describing the temporal relationships between affective disorders, anxiety disorders, psychosis and substance abuse disorders.

The original instrument was developed in five sites in a sample of 81 subjects. Test-retest reliability (with a 4–10 day interval) was high for DSM-III-R diagnoses of major depression, bipolar disorder and schizophrenia, with kappa coefficients of 0.94, 0.96 and 0.75, respectively. A lower kappa coefficient (kappa = 0.31) was obtained for schizoaffective disorder.

In 1999 French version of DIGS interview was tested in Swiss population [14] through a study of 136 cases with the referent diagnoses used in the original study. The French group also carried out a study to test the interview in alcohol/drug abuse and dependence, together with antisocial personality disorder [1]. They obtained good results in studies of family aggregation and comorbidity with psychiatric disorders, including substance abuse and antisocial personality disorders.

Other translations have reported a good reliability of the DIGS in Hindi population with 20 patients [4] and more recently in Korean population [5] in a sample of 53 patients. DIGS was also validated in the Colombian population [13] with a sample of 65 patients and with kappa coefficients for schizophrenia of 0.87, 1.0 for bipolar disorder, 0.92 for major depression and for schizoaffective disorder of 0.84.

The purpose of the present study was to assess the reliability and validity of DSM-IV diagnoses obtained by the Spanish version of the DIGS. The DIGS interview was tested by four groups, three of them working in a Research Network on Genotyping and Mental Illness.

2. Subject and method

2.1. Sample

The reliability of diagnoses for mood disorders and psychosis, obtained through the Spanish population version of the DIGS, was assessed in 95 patients from outpatient and inpatient facilities of four Spanish general hospitals: Hospital Joan March, Hospital Universitario Son Dureta, Hospital del Mar y Hospital Ramón y Cajal. The patients had DSM-IV diagnoses of schizophrenia, bipolar disorder, major depression and schizoaffective disorder. Four patients were not included in the final sample due to insufficient data. One patient from the initial sample did not participate in the re-test phase as he failed to attend the scheduled meeting.

Patients were asked to sign an informed consent form. This research was approved by the ethics committee of the Balearic Islands health administration.

2.2. Procedure

Translation process. The 3.0 English version of DIGS was used. The Spanish version of the interview was created in several phases by a team of psychiatrists and psychologists from the Balearic Islands University (Palma de Mallorca, Spain). It was initially translated into Spanish and then two psychiatrists and one psychologist applied the interview in a pilot study to a small population, in order to verify the interpretation of the questions and modify some terms to improve their comprehension. The resulting document was then back translated and compared with the original version of DIGS, which lead to minor changes for the final version.

The Spanish version of DIGS was structured similar to the French version, and incorporated some of the modifications added to the latter:

1. An optional screening question for mania was added in order to lower the threshold for entering this section by asking whether there was objective evidence of elated mood (friends or family members have observed that the subject’s mood was higher than normal).
2. Optional questions were added to allow a better temporal assessment of the last episode in both the major depression and mania sections.

Standardization process. A team of interviewers (psychiatrists and psychologists) with experience in clinical interviews carried out the project in the aforementioned centers. A researcher from the French group who has extensive experience with the interview (M. Preisig) worked for one week with the entire Spanish group in training sessions utilizing the DIGS manual in Spanish and real patient cases from the different centers.

Patients were selected from the inpatient and outpatient units of the noted sites. All interviewers were blinded as to the recruitment source (specific clinical unit) and referral diagnosis of subjects (patients in habitual monitoring and diagnostic procedures to guarantee DSM-IV inclusion criteria).

Patients were interviewed using the DIGS by a member of the team in the presence of an observer (co-rater). Simultaneously and independently the observer completed another DIGS. At the end of the interview, the interviewer and observer independently assigned the definitive diagnosis based on the DSM-IV. The average duration of each interview was 60 min.

Between four and six weeks later, subjects were contacted again for the retest interview. The retest was conducted by a third team member who had no information regarding the first interview (was blinded as to the diagnosis and origin of the patient as well as two previous DIGS interviews). The time interval (4–6 weeks) between test and retest was longer than that used to test the original instrument (4 to 14 days). The study coordinators felt that this longer interval would help guarantee the diagnostic stability.

The Longitudinal Expert All Data (LEAD) procedure has been proposed as a criterion for the assessment of the procedural validity of diagnostic instruments [9,16]. Each patient attending psychiatrist served as the expert for that patient. LEAD experts gave at the end of the study the DSM-IV diagnoses of each patient in two times frames: current (within the
past year, which was used for the study) and past (before previous 12 months). To develop the LEAD diagnoses the experts used all available material: clinical interviews, evaluations by nurses staff, medical consultants, laboratory results, previous medical records of treatments in both inpatients and outpatients units, data provided from family of the patient, when possible. At the end of the procedure the expert had to fill in a questionnaire the presence or absence of DSM-IV diagnoses. The LEAD experts were blind to the DIGS evaluations.

3. Data analysis

For the analysis of inter-rater and test-retest reliability 5 × 5 tables were constructed based on the diagnostic categories of schizophrenia, bipolar disorder (type I), major depression, schizoaffective disorder and other diagnosis. Consistency of diagnosis was calculated as Cohen’s Kappa coefficient [3]. We also compared the diagnoses based on the DIGS with the LEAD procedure. When one patient had more than one diagnosis only one was taken into account to calculate the Kappa index. Overall Kappa coefficient was calculated from these 5 × 5 tables and individual kappa coefficients were calculated from 2 × 2 tables that were constructed for each disorder based on the presence or absence of the specific diagnostic. According to Fleiss [7], the agreement is considered as “excellent” when the kappas are over 0.75, “fair” to “good” when the kappas are between 0.40 and 0.74, and “poor” when they are down of 0.40 [17]. Kappa statistics with 95% confidence intervals and Yule statistics were analyzed with the SPSS 12.00 statistical software package (Chicago, SPSS, Inc).

4. Results

4.1. Sample

Forty-eight% of the patients were men and 52% women. The average age of subjects was 43.5 years old (19–68). 94.7% of patients interviewed also participated in the retest phase (90 of 95 patients).

4.2. Inter-rater reliability

The overall kappa coefficient was 0.956. The kappa value for individual diagnosis varied from major depression (kappa = 0.877) to schizophrenia (kappa = 1), all with excellent kappa: schizophrenia (kappa = 1), schizoaffective disorder (kappa = 0.1), bipolar type I disorder (kappa = 0.903), major depression (kappa = 0.877). Results are provided in Table 1.

4.3. Test-retest reliability

The frequencies and results of the test-retest analysis are presented in Table 2. The computed overall kappa of 0.926 is similar to that of inter-rater analysis. With the exception of major depressive disorder (kappa = 0.776), the kappas for all individual target diagnoses ranged from of 0.846 to 1, which is excellent agreement.

4.4. Validity: comparison between diagnoses obtained by DIGS (interviewer rating) and LEAD procedure

Table 3 shows the consistency of DSM-IV diagnoses with LEAD procedure. Overall concurrent validity was excellent.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Inter-rater reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td>Bipolar I</td>
</tr>
<tr>
<td>Bipolar I</td>
<td>28</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>0</td>
</tr>
<tr>
<td>Schizoaffective disorder</td>
<td>0</td>
</tr>
<tr>
<td>Major depression</td>
<td>3</td>
</tr>
<tr>
<td>Other disorders</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Test-retest reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td>Bipolar I</td>
</tr>
<tr>
<td>Bipolar I</td>
<td>28</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>0</td>
</tr>
<tr>
<td>Schizoaffective disorder</td>
<td>0</td>
</tr>
<tr>
<td>Major depression</td>
<td>3</td>
</tr>
<tr>
<td>Other disorders</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
</tr>
</tbody>
</table>
(kappa = 0.797), the kappa values for schizophrenia, bipolar disorder (type I) and schizoaffective disorder were also in the excellent range (0.756 to 0.825). However, concurrent validity was just fair to good (kappa = 0.704) for major depression.

5. Discussion

Comparing the validations of different countries is controversial but these results show a high level of agreement than those obtained in previous validation studies and even with those in the original North American version [6,15,11,19]. The French version [12] obtained a lowest kappa for schizoaffective disorder (0.60 to 0.87) than for the other disorders (0.85 to 1.00) as well as a test-retest reliability of 0.38 to 0.48 compared to 0.62 to 0.65 for the rest of the diagnoses. The results are similar in the Colombian population study by Palacio et al. [13], with “excellent” reliability intervals for schizophrenia, bipolar disorder and major depression, “good” for “other diagnoses” and “poor” for schizoaffective patients (see Table 4).

In our study these coefficients are also “good” for schizoaffective disorder (0.846). It has been argued that lower kappas in this disorder could be due to the complexity of clinical data collection, the duration of affective and psychotic symptoms and to overlapping [12,14]. In some cases this result was due to a low prevalence of the disorder in the selected sample [14].

The agreements for major depression are lower than in the original study and the Colombian [13] and Korean [5] validations. The French study, with a kappa of 0.59, also offers results similar to the Spanish population. Major depression is apparently homogeneous across all reliability and validity studies.

The DIGS was not designed for routine clinical use, but rather for genetic research and it must be considered as such. The administration time is an important issue at clinical level and research setting. Given the duration of the interview, some patients may have tried to shorten it by offering negative responses (especially in the retest phase) as occurs with many other psychometric instruments. Nevertheless, the test-retest level of concordance for the average administration time was high. Depending on the objective of the study it is possible to use separately sections (i.e.: psychotic or alcoholism section).

The present study address DIGS validity using the LEAD procedure as “gold standard”. The overall reliability of the LEAD procedure is comparable to other diagnostic methods and has been used to validate other structured interviews in psychiatry [18].

The size of our sample might be not enough in some diagnoses, for this reason we presented also the Yule together with the kappa, but these diagnoses should be retested in samples with higher prevalences.

It is still unclear whether the DIGS can increase the potential to develop quantitative phenotypes. The habitual categorical phenotypes permit partial advancing in psychiatric research. Using DIGS may allow analyzing multiple genetic studies and extracting more reliable results comparisons.

### Table 3

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Interview/LEAD</th>
<th>Bipolar I</th>
<th>Schizophrenia</th>
<th>Schizoaffective disorder</th>
<th>Major depression</th>
<th>Other disorders</th>
<th>Total</th>
<th>Kappa</th>
<th>95% C.I.</th>
<th>Yule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bipolar I</td>
<td>25</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>0.825</td>
<td>0.889–0.761</td>
<td>0.754</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>13</td>
<td>0.797</td>
<td>0.884–0.71</td>
<td>0.655</td>
<td></td>
</tr>
<tr>
<td>Schizoaffective disorder</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0.756</td>
<td>0.872–0.64</td>
<td>0.556</td>
<td></td>
</tr>
<tr>
<td>Major depression</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>27</td>
<td>0</td>
<td>30</td>
<td>0.704</td>
<td>0.783–0.625</td>
<td>0.586</td>
<td></td>
</tr>
<tr>
<td>Other disorders</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>12</td>
<td>14</td>
<td>0.810</td>
<td>0.901–0.719</td>
<td>0.667</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>13</td>
<td>7</td>
<td>30</td>
<td>12</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
<td>0.797</td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>81</td>
<td>99</td>
<td>65</td>
<td>53</td>
<td>20</td>
<td>95</td>
</tr>
<tr>
<td>Instrument</td>
<td>DIGS</td>
<td>DIGS</td>
<td>DIGS</td>
<td>DIGS-K</td>
<td>DIGS</td>
<td>DIGS</td>
</tr>
<tr>
<td>Interval</td>
<td>4–10 days</td>
<td>42 days</td>
<td>42 days</td>
<td>1–50 weeks</td>
<td>1 week</td>
<td>28–42 days</td>
</tr>
<tr>
<td>Diagnostic criteria</td>
<td>DSM-III-R</td>
<td>DSM-IV</td>
<td>DSM-IV</td>
<td>DSM-IV</td>
<td>DSM-IV</td>
<td>DSM-IV</td>
</tr>
</tbody>
</table>

### Diagnoses

- **Bipolar Dis.** 0.96, 0.63, 1, 0.74, --, 0.903
- **Schizophrenia** 0.75, 0.72, 0.87, 1, --, 1
- **Schizoaffective dis.** 0.31, 0.40, 0.84, 0.43, --, 0.846
- **Major depression** 0.94, 0.59, 0.92, 1, --, 0.776
- **Other disorders** --, --, 0.65, --, --, 1
- **No disorder** 0.86, 0.65, 0.88, --, --, --
6. Conclusion

The DIGS interview proves to be a reliable and valid instrument for genetic studies in the Spanish population. It could even be used as a research instrument in other biological marker studies, given its diagnostic capacity for both the current and previous episodes. The main interest in testing the instrument in several countries lies in its potential to compare results with previous studies utilizing criteria based on a non-DSM-IV classifying system, as Nurnberger et al. [12] suggested in his initial description. Having this instrument available and validated in the Spanish population seemed clearly advantageous for: using at the very heart of the aforementioned Network and for other Spanish groups working in the field; replicating previous genetics studies; comparing diagnosis across countries; and pooling samples to increase statistical power.

The Spanish validated version will be at disposal on the web page of the Spanish Research Network on Genotyping and Psychiatric Genetics (http://www.rgpg.net).

Acknowledgments

The Spanish Research Network on Genotyping and Psychiatric Genetics is supported by the Instituto Carlos III (Ministry of Health): GO3/184. This study was financed in part by an unrestricted grant from Boehringer-Ingelheim of Spain. Our special thanks goes to doctor Martin Preisig, the researcher who supervised the training of the Spanish interviewers.

References