Berres M, Monsch AU, Bernasconi F, Thalmann B, Stahelin H (2000)Normal ranges of neuropsychological tests for the diagnosis of Alzheimer's disease. *Stud Health Technol Inform*.;77:195-9.

Normal ranges of neuropsychological tests for the diagnosis of Alzheimer's disease

Manfred Berres, *Novartis Pharma AG, WSJ-27.5.016, CH-4002 Basel*
Andreas U. Monsch, Florence Bernasconi, Beat Thalmann, Hannes B. Stähelin, *Geriatric University Hospital, Kantonsspital, Hebelstrasse 10, CH-4031 Basel*

The diagnosis of early stage dementia is a highly complex process involving not only a somatic examination but also a neuropsychological assessment of the patient's cognitive capability. The American 'Consortium to Establish a Registry for Alzheimer's Disease' (CERAD) has proposed a set of tests in English which has been translated into German. This paper presents the statistical methodology applied to determine normal ranges adjusted for demographic variables for the German CERAD neuropsychological assessment battery (CERAD-NAB).

The study population consists of participants of the Basel Study on the Elderly (Project BASEL) which aims at identifying preclinical markers of Alzheimer's disease. The normative sample has been defined by carefully excluding potentially relevant medical history and concomitant diseases and consists of 617 participants which are between 53 and 92 years old. Test results should be adjusted for gender, age, and years of education. For this purpose, a set of linear models including these predictors and subsets of their interactions and squares was evaluated for all 11 test scores derived from the CERAD-NAB battery. Model selection was based on the *PRESS* (predicted residual sum of squares) statistic. Although a strict application of this criterion selected 6 different models, a slight compromise allowed to fit all test scores by two models.

In several tests of the CERAD-NAB many participants achieve maximal scores. Residuals of such test scores are heavily skewed. An arcsine transformation has been tuned to the data, so that residuals are close to a normal distribution, at least for residuals in the lower quartile which is relevant in diagnosing cognitive impairment. Test results are finally presented as *z-scores* which can be easily compared to a standard normal distribution.

The evaluation of the CERAD-NAB is implemented on the Internet and in an Excel application.

1. Introduction

Subjective complaints about cognitive deficiencies become more frequent, as more people are getting older. While in some cases these complaints can be attributed to natural worsening of memory with age, other cases request a differential diagnostic process between for example Alzheimer's disease, depression and cerebro-vascular disease. Neuropsychological tests play a central role in assessing the cognitive ability of individuals. A frequently used assessment battery for English speaking patients was proposed by the American 'Consortium to Establish a Registry for Alzheimer's Disease' (CERAD-NAB) [1]. This battery consists of 5 tests: *Verbal Fluency: Animal Category*, a short form of *the Boston Naming Test*, *Mini-Mental State Examination* (*MMSE*), *Verbal Memory Test* consisting of word list learning, delayed recall, and a recognition procedure and *Constructional Pratice* (including delayed recall). These five tests are administered in eight steps and 11 scores are calculated. Using such tests as diagnostic tools is based on two prerequisites: (1) normal ranges should have been defined and (2) the test should be able to discriminate between healthy individuals and patients in an early stage of dementia.

This paper deals with the definition of normal ranges for a German translation of the CERAD-NAB battery. Different approaches have been used, the most simple one is to stratify a sample of normal subjects (e.g. by gender, age and education) and define the 5th percentile as lower limit of the normal range, if diseased individuals achieve low scores. More information is gained, if the score is given on a normal scale, e.g. a standard-normal *z-score* or, equivalently, an *IQ-score* with a mean of 100 and a standard deviation of 15. The 5th percentile of the *z-score* is –1.645 and the 5th percentile of the IQ score is 75.3. However, (raw) test scores often do not exhibit a normal distribution. They are often transformed to empirical normal distributions by taking the quantile of the cumulative distribution function of the raw test scores. This has been proposed long ago by McCall [2] and is still recommended [e.g. 3]. It is a highly data-driven procedure, because a separate transformation is derived for each stratum. If the sample sizes of the strata are small, it may even end up with non-monotonic changes of the lower limit with age where a monotone relation between age and the normal range is expected.

Such inconsistencies can be avoided by setting up a regression model for the dependence of the raw scores on the demographic variables. The residuals of the regression model are equivalent to raw scores adjusted for the predictors. If the residuals deviate from a normal distribution and the relation in the model is not linear, both problems can often be cured by an intuitively reasonable non-linear transformation of the raw scores. Dividing the residuals by their standard deviation yields *z-scores* of the raw test values. These *z-scores* are defined in a consistent way for the whole normative sample. Instead of a collection of tables mapping raw scores to standard scores, we arrive at rather simple formulas which can be easily implemented in computer programs. In addition, the *z-scores* can be interpreted easily by any experienced clinician.

2. Material and Methods

The population for this normative study is derived from the *Basel Study on the Elderly* (*Project BASEL*, [4]). The population is a subset of the *Basel Longitudinal Study* [5]. *Project BASEL* is a prospective study and aims at identifying preclinical markers of Alzheimer's disease. Exclusion criteria have been based on medical history as well as present clinical findings. Participants were excluded, if they suffered from CNS related, psychiatric or severe systemic diseases, had experiences like unconsciousness for more than 5 minutes, brain surgery or general anesthesia within 3 months, took toxic substances or more than 80 g of alcohol per day, had auditive, visual, lingual, sensory or motor activity deficits or continuous pain (cf. [6]). All individuals for the normative sample were native German-speaking.

Based on these criteria 617 participants of the *Project BASEL* study were eligible for the normative sample, 432 males and 185 females. Their age ranged from 53 to 92 years with a mean of 70.3 years (SD: 7.6 years). Education is a relevant predictor in many cognitive tests, it was measured in years (at school, university, vocational training) and ranged from 6 to 20 years with a mean of 12.7 years (SD: 3.1 years). Only 3 participants had less than 8 years of education.

The CERAD-NAB was considered the most appropriate diagnostic tool to screen participants of *Project BASEL* for cognitive impairment. With permission from CERAD at Duke University in Durham, NC, USA, we translated the American CERAD-NAB into German. Normal ranges of this new German version needed to be defined.

Normal ranges should be adjusted for age, years of education and gender. Since regression models were to be applied for this purpose, the first step was to select an appropriate model. The following sets of predictors were considered to cover a sufficiently

broad range of models for this selection (squared terms and age by education interaction are centered by the median per gender-category):

```
AGE EDUCATION
AGE EDUCATION AGESQ[1]
AGE EDUCATION EDUCATIONSQ[2]
AGE EDUCATION AGE*EDUCATION[3]
AGE EDUCATION GENDER
AGE EDUCATION GENDER AGESQ
AGE EDUCATION GENDER EDUCATIONSQ
AGE EDUCATION GENDER AGE*EDUCATION
AGE EDUCATION GENDER AGE*GENDER
AGE EDUCATION GENDER EDUCATION*GENDER
AGE EDUCATION GENDER AGE*GENDER EDUCATION*GENDER
```

The *PRESS* statistic was used as criterion for model selection. The *PRESS* statistic is the sum of squares of the predicted residuals for all observations where the *predicted residual* is the residual for an observation that results from dropping this observation from the model estimation. Each predicted residual is thus pretending to predict a 'new' observation and this idea is well suited for models that will be applied to future observations in practice.

The residuals from several tests of the CERAD-NAB are far from being normally distributed. This is in most cases due to a ceiling effect, which cuts off the score of very apt people at the maximum. It is particularly true for the *Mini-Mental State Examination* (*MMSE*), which checks basic abilities like knowing today's date, repeating single words, carrying out a 3-step command. Most healthy people achieve a score of 28 or more and a substantial portion achieves the maximal score of 30. Translating the score to the proportion of items solved, most healthy people achieve a proportion close to 1. A well-known transformation for that type of data is the arcsine transformation. It is usually applied to the square root of a proportion ($\arcsin(\sqrt{x}), x \in [0,1]$) resulting in a symmetric function with respect to x=1/2. For the present data only the transformation near x=1 is needed, hence we drop the square root. But we introduce a tuning parameter that allows to adapt the transformation to achieve residuals that follow the normal distribution law at least at the left hand side, because only low residuals are relevant for differentiating between healthy and diseased subjects. For the *MMSE* the class of functions is

$$y = \arcsin\left(\frac{MMSE}{30 + a}\right),\ MMSE \in \{0,1,\ldots,30\},\ a \geq 0\,. \qquad (*)$$

The model selection should be confirmed after transformations. Finally, the residuals are divided by their standard deviation and displayed in a Q-Q normal probability plot.

The *z-score* of a new individual is determined as

$$z = \frac{transform(raw\ score) - prediction}{s_e},$$

where *prediction* is calculated from the parameters of the respective regression model and s_e (= *Root MSE*) is the standard deviation of the residuals in this model.

3. Results

Model selection for the whole CERAD-NAB was governed by a compromise of minimizing the *PRESS* statistic and selecting only a few different models. A strict

[1] AGESQ = $(AGE-71)^2$ for males and = $(AGE-70)^2$ for females

[2] EDUCATIONSQ = $(EDUCATION-12)^2$ for males and = $(EDUCATION-11)^2$ for females

[3] AGE*EDUCATION= (EDUCATION-12)(AGE-71) for males and = (EDUCATION-11)(AGE-70) for females

minimization of the *PRESS* statistic would have selected 6 different models. One of the tests (not shown here) needed an extra quadratic term for EDUCATION to fit the 3 aforementioned participants with less than 8 years of education. Removing these 3 individuals from the sample the quadratic term could be dropped. The normal ranges were finally based on 614 participants and are only applicable for individuals with 8 or more years of education. With the stress criterion slightly relaxed, all scores could be modeled by either the main effects age, education and gender or by adding the interaction of education and gender to the main effects.

Results on *verbal fluency* and on the *MMSE* are presented in more detail.

Verbal fluency did not need a transformation. The model with interaction achieved the minimal *PRESS* value of 16286 ($N = 614$), the other *PRESS* values ranging between 16302 and16717. Parameter estimates are shown in Table 1, the quantile plot is presented in the left panel of Figure 1.

The *PRESS* statistic for the raw *Mini-Mental State Examination Score (MMSE)* was minimized by the model of main effects plus age by gender interaction, which was negligibly better than the main effects model. *MMSE* was transformed by the arcsine function (*) with parameter a=0.1, which was determined by a grid search to optimize the left-hand tail of the distribution of residuals in the main effects model. The minimum of the *PRESS* statistic for the transformed score was achieved with the same model as before. Its value of 10.99 was again close to that of the main-effects model (11.04) and this latter model was finally chosen. Parameter estimates for the transformed scores are reported in Table 1; the right panel of Figure 1 shows the quantile plot.

Table 1: Parameters (standard errors in parentheses), R-square and standard deviation of residuals for *Verbal fluency* and transformed *MMSE* (Education is **not** centered in the interaction term, $N = 614$).

Test score	Intercept	Age	Education	Gender	Education*Gender	
Verbal fluency	30.34 (2.16)	-0.1825 (0.0276)	0.288 (0.082)	-6.87 (1.94)	0.642 (0.158)	R^2=0.141 s_e=5.13
MMSE	1.5723 (0.0550)	-0.00562 (0.00072)	0.01059 (0.00182)	0.0369 (0.0124)	Not included	R^2=0.132 s_e=0.1337

4. Discussion

Psychologists often use the normal quantiles of the empirical cumulative distribution to define normal scores. This is usually repeated for different strata of the normative sample. The method proposed here yields more consistent results, because it is based on a uniform model for the whole sample. The approach is thereby more appealing. It has been used by [7]. While Gladsjo et al. [7] applied stepwise regression and performed a cross-validation, the present method selects the optimal model by minimizing the *PRESS* statistic. This includes a very thorough form of cross-validating the model for each observation and avoids the need to sacrifice a substantial percentage of the sample for validation.

It is well-known that the *PRESS* statistic is high due to lack of fit, if the model contains too few terms and is also high, if the model contains too many terms. Between these cases it assumes its minimal value for the model with optimal predictive power.

In regression models normal ranges, also called *tolerance intervals* [8], usually have the form of a hyperbola, accounting for increasing standard errors from the center of the predictors to their boundary values. This increase is less than 1% of the difference between a fitted value and its lower limit for the main-effects model, and less than 2.8% of this dif-

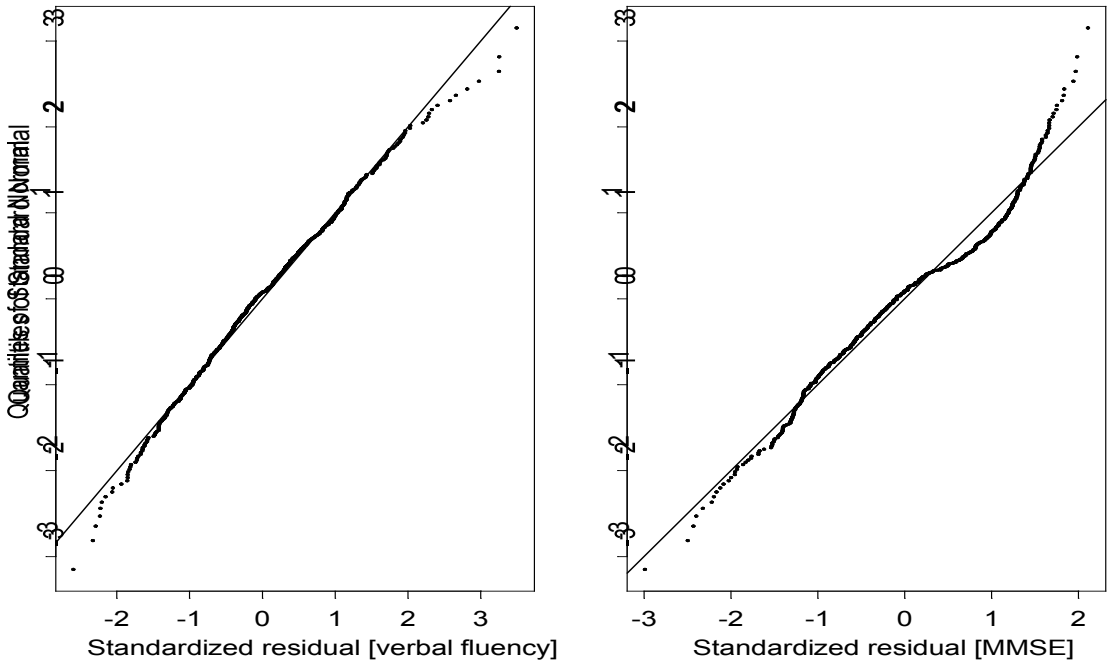

Figure 1: Q-Q normal probability plots of the residuals for *verbal fluency* (left) and transformed *MMSE* (right) in their finally selected models. Approximation by the normal distribution is only needed for negative residuals.

ference in the model containing an interaction term. Considering the precision of the data, this could be neglected in the present sample.

When diagnosing new individuals, only negative *z*-scores need closer consideration. Any deviation from the normal distribution for positive *z*-scores - as it is seen with the *MMSE* - does not invalidate the diagnosis.

To aid clinicians in judging the quality of the data, three diagnostic *z-scores* can be calculated, one for the individual's actual raw score, one for the next higher and one for the next lower raw score. This should protect from misinterpreting accidental low test scores.

The calculation of *z*-scores for the complete CERAD-NAB including the evaluation at neighboring raw scores is implemented in an Excel® application, which can be purchased from the second author (A.U.M.). A similar application is implemented on the internet. Professionals can register on http://www.healthandage.com/ (→ Physicians and Reasearchers → Alzheimer's → Geriatric Assessment).

References

[1] J. C. Morris, A. Heyman, R. C. Mohs, J. P. Hughes, G. van Belle, G. Fillenbaum, E. D. Mellits and C. Clark, The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease, *Neurology* **39,** (1989) 1159-1165.

[2] W. A. McCall, Measurement. New York, 1939.

[3] M. Röhr, H. Lohse and R. Ludwig, Statistische Verfahren. Verlag Harri Deutsch, Thun, 1983.

[4] H.B. Stähelin and M. T. Widmer, 39 Jahre Basler-Studie (1959-1998), *Geriatrie Praxis* **4** (1998) 34-40.

[5] L. K. Widmer, H.B. Stähelin, C. Nissen and A. da Silva (ed.), Venen-, Arterien-Krankheiten, koronare Herzkrankheit bei Berufstätigen. Verlag Hans Huber, Bern, 1981.

[6] Alzheimer Forum Schweiz, Diagnostik und Therapie der Alzheimer Krankheit: Ein Konsensus für die Schweiz, *Schweizerische Aerztezeitung* **80 (**1999) 14ff.

[7] J. A. Gladsjo, S. Walden Miller and R. K. Heaton, Norms for Letter and Category Fluency: Demografic Corrections for Age, Education, and Ethnicity. Psychological Assessment Resources, Inc. 1999.

[8] P. Armitage and T. Colton (Ed), Encyclopedia of Biostatistics. John Wiley & Sons, Chichester, 1998.

This study was supported by a grant from the Swiss National Science Foundation (NF 3200-049107)