Statistical advances in clinical neuropsychology
Agelink van Rentergem Zandvliet, J.A.

Citation for published version (APA):
Consider the case of an elderly man, Albert, who has fallen off his bicycle because of a reckless scooter driver, and is taken to hospital. Albert’s physical injuries do not seem to be severe, and he is soon released. After some weeks, Albert’s wife contacts the hospital, because she thinks his memory has worsened. Albert himself is not aware of any changes. A neurologist sends Albert to a clinical neuropsychologist for a neuropsychological assessment. The clinical neuropsychologist has to decide between several options. Is Albert’s memory indeed bad, and is this consistent with a traumatic brain injury from his accident? Is Albert’s wife perhaps overly worried, and is Albert’s memory consistent with what would be expected for a man of his age? Or is Albert’s memory bad, and is this part of a larger problem, perhaps a disorder like Alzheimer’s disease?

It is important to Albert, his wife, and the hospital that the neuropsychological assessment is as reliable as possible. If Albert’s wife is indeed overly worried, this should be discovered, so these worries can be resolved. If Albert is suffering from a traumatic brain injury, this should be discovered so the hospital can further investigate this injury (Maas, Stocchetti, & Bullock, 2008). If Albert is suffering from a disorder like Alzheimer’s disease, this should be discovered so Albert can start with treatment (Small et al., 1997). It is therefore crucial that the neuropsychological assessment is successful in providing clarity to all parties.

Outside clinical practice, neuropsychological assessments are also performed in research. This may be done in studies that evaluate a new treatment, for example, to ameliorate the symptoms of dementia. Neuropsychological assessments are also used to detect adverse effects of treatments on cognitive functioning. Cognitive functioning may be affected by a wide variety of pharmaceutical treatments, for example psychiatric drugs (Moore & O’Keeffe, 1999) or drugs that are aimed at a different target entirely, like chemotherapy (de Ruiter et al., 2011), and non-pharmaceutical treatments such as deep brain stimulation (Smeding, Speelman, Huizenga, Schuurman, & Schmand, 2006), or surgery of the brain (Spencer & Huh, 2008) or heart (Selnes et al., 2012). In studying these treatments, it is important that the neuropsychological assessment is highly reliable, because otherwise, harmful side effects may be overlooked.

Studies may also use neuropsychological assessments to evaluate whether cognitive functions are impaired in a particular disorder, be it a disorder of the brain like schizophrenia (Schaefer, Giangrande,
Weinberger, & Dickinson, 2013), or a disorder that may indirectly affect cognition, like liver cirrhosis (O’Carroll et al., 1991) or diabetes (Cheng, Huang, Deng, & Wang, 2013). If this is the case, researchers may study what characteristics of the patients predict which patients are affected, as some cognitive problems for example primarily occur in older patients. The reliability of the neuropsychological assessment is again critical in identifying those with cognitive impairment, and those without.

The goal of this thesis is to improve the reliability of neuropsychological assessment, specifically by improving the normative comparison procedure. This thesis is embedded in the Advanced Neuropsychological Diagnostics Infrastructure (ANDI) project. This thesis discusses multiple statistical methods that were developed for the ANDI project to improve normative comparisons. In this chapter, several key concepts are introduced, and the specific goals for the ANDI project are outlined. Then, an overview of the remaining chapters is given.

1.1 WHAT IS NEUROPSYCHOLOGICAL ASSESSMENT?

In clinical neuropsychology, patients are assessed to characterize their cognitive functioning. Subjective cognitive complaints, an accident or stroke, or a disorder like Parkinson’s disease are all indications that cognitive functioning may be impaired, and can thus be reasons for a neuropsychological assessment (Lezak, Howieson, Bigler, & Tranel, 2012). This type of assessment is standardized, in order to make the results comparable between different clinicians and patients. Therefore, standardized neuropsychological tasks are used, which may consist of memorizing a message, naming objects in pictures, enumerating as many words starting with a particular letter as possible within one minute, or tracing a pattern with a pencil (Strauss, Sherman, & Spreen, 2006). Each of these tests is designed to tap into a different part of cognitive functioning, such as memory, psychomotor skills or attention. By measuring these cognitive functions, the neuropsychologist can decide whether a patient’s cognition is impaired. The goal of the ANDI project and this thesis is to improve the precision with which this decision is made.

1.1.1 Normative comparisons

To decide whether a particular score on a test is indicative of impairment, a certain reference standard has to be used. For almost all neuropsychological tasks, there is no score that can be considered indicative of impairment in an absolute sense. Rather, patients’ test scores are considered relative to those obtained by a group of healthy people (Crawford & Garthwaite, 2002), typically called a norm group or normative sample. If a patient’s test score is lower than those obtained by
1.1 WHAT IS NEUROPSYCHOLOGICAL ASSESSMENT?

the majority of healthy people, this is an indication for impairment. To be able to make such a judgment, called a normative comparison, data from many healthy participants who have completed neuropsychological tests are needed. Therefore, one of the goals of the ANDI project is to establish a large normative dataset, to improve normative comparisons.

1.1.2 Multivariate normative comparisons

The idea for the ANDI project came in part from the introduction of a new statistical technique for normative comparisons, called multivariate normative comparisons. Traditionally, normative comparisons are performed for a single neuropsychological test at a time, and are therefore univariate (Huizenga, Smeding, Grasman, & Schmand, 2007). This univariate approach has two disadvantages. The first is that it does not match clinical intuition, as results on tests are not interpreted in isolation by clinicians, but are interpreted in the light of results on other tests. For example, a low score on two delayed memory tests is interpreted differently when found in a patient with high scores on all other tests, than when found in a patient with low scores on all tests.

The second disadvantage is an increased number of times that a patient is incorrectly classified as cognitively impaired, i.e., that the assessment indicates impairment, while the patient is in fact not cognitively impaired. The aim is to keep the number of persons that are mislabeled like this low. However, for each normative comparison, there is a probability that this comparison will by chance indicate impairment, which is called a false positive result. This is the case even for a cognitively healthy person. With univariate normative comparisons, a comparison is performed for every neuropsychological test score, and the probability of at least one false positive result for a healthy person becomes larger and larger by chance when additional normative comparisons continue to be made. For example, a healthy person has a higher chance of a false positive result if this person is given many opportunities, in tests of verbal memory, executive functions, motor speed, attention, naming, and fluency. This risk is lower if only a single test is administered. There is no good way of knowing for a new patient whether a finding of cognitive impairment is incorrect or not, and if no steps are taken to control the number of times that incorrect classifications are made, many healthy people may inadvertently be labeled as cognitively impaired by univariate comparisons (Binder, Iverson, & Brooks, 2009).

These disadvantages are not found in multivariate normative comparisons. First, multivariate normative comparisons analyze the entire profile of test scores, similarly to how a clinician takes into account the whole profile of scores (Huizenga et al., 2007). This means
that the analysis takes into account whether the patient’s profile of scores is common in healthy participants. For example, the combination of a very high score on immediate recall of words and a low score on an attention test is common among healthy people. The combination of a very high score on immediate recall of words and a low score on recall of words after 30 minutes is something that is not observed in healthy people. This profile of scores could indicate impairment of memory storage.

Second, multivariate normative comparisons always provide a single comparison. This means that if a profile of twenty-five test scores is tested in a normative comparison, this entails a single comparison, just like a profile of five test scores would. Because there is only a single comparison, the probability of finding a false positive result, and thus incorrectly classifying a cognitively healthy person as cognitively impaired, is under control, no matter how many neuropsychological test scores are entered into the comparison.

One problem for multivariate normative comparisons is that it requires that the healthy people in the normative group have completed multiple tests. Ideally, they would have completed all the same neuropsychological tests that the patient completes in the assessment. This type of normative data is not available. Therefore, one of the goals of the ANDI project and this thesis is to provide normative data from healthy participants who have completed multiple tests, in order to facilitate the implementation of multivariate normative comparisons.

1.1.3 Demographic corrections

Another important aspect where normative comparisons can be improved is the area of demographic corrections. When evaluating a patient’s scores for the presence of a cognitive impairment due to a disorder, the cognitive impairment is best detected when the healthy participants in the normative group are similar to the patient in characteristics unrelated to the disorder. What this means is that a neuropsychological assessment for a 72-year-old patient is most reliable when we compare his or her scores to those obtained by healthy 72-year-olds. Such corrections are commonly performed for age. However, level of education also predicts cognitive test scores. Therefore, we ideally compare test scores from patients with low education to those obtained by healthy people with low education, to increase sensitivity. Sex generally plays a smaller role, but there may be a small increase in sensitivity if male patients are compared to healthy men, and female patients are compared to healthy women (Lezak et al., 2012). Which demographic variables to correct for depends on the type of test used (de Vent, Agelink van Rentergem, Murre, ANDI Consortium, & Huizenga, 2016a, this thesis).
The available normative data for neuropsychological tests rarely allow demographic correction for age, sex, and level of education. Instead, normative data may be available for different ages, but not for different levels of education and sexes. Also, because demographic corrections require many different participants, data may not be available for individual ages. This means that, for example, a 72-year-old patient has to be compared to a group of 70 to 80-year-olds, which decreases sensitivity (Testa, Winicki, Pearlson, Gordon, & Schretlen, 2009). Therefore, one of the goals of the ANDI project and this thesis is to provide normative data from large numbers of healthy participants who have completed neuropsychological tests and for whom age, sex, and level of education is known, in order to facilitate more precise demographic corrections.

1.1.4 **Online availability**

A third major theme of this thesis and the ANDI project is using internet-based technology to aid clinical neuropsychology. Normative comparisons for a single test that are corrected for age are typically performed by looking up the patient’s score in a printed table of age bins and scores. Scores for different ages, sexes, and levels of education become more difficult to tabulate and to look up. The same is true for multivariate normative comparisons: Multivariate normative comparisons cannot be easily performed with printed tables, as there are many dimensions if there are multiple tests involved. One solution is to no longer look up the results by hand, but to let computers calculate the results (Miller & Barr, 2017). Therefore, one goal of the ANDI project is to build a website on which clinicians can perform normative comparisons of their patient data. This allows clinicians to use these statistically sophisticated techniques anywhere, and allows us to update the procedures as new data and methods become available.

Another advantage of using the internet is that it becomes easy to share information with a large number of clinicians and scientists. The normative comparison procedures described in this thesis are in principle not restricted to the field of clinical neuropsychology. Therefore, one could take the software and apply it in other fields of psychology or in other disciplines, such as medicine. To facilitate this, the computer code for the methods developed in this thesis and that are used in the ANDI project are freely available online. A second advantage of sharing the code of the ANDI project is transparency (Poldrack & Gorgolewski, 2014). This means that the implementation of the methods described in this thesis are also available to any user or programmer who wants to review and criticize the method (Nosek et al., 2015).
1.2 OVERVIEW OF THIS THESIS

In the second chapter, a method is described for establishing a normative dataset that fits the goals outlined above. This method is based on the combination of data from healthy people who have already taken part in research, for example as a participant in a control group in a clinical study, or as a participant in an epidemiological community study. By combining data from multiple studies, it becomes possible to obtain large numbers of participants, who are demographically diverse, and have completed many different tests. This chapter explains standardized procedures for removing outlying values, determining what demographic variables to use in corrections, and finding appropriate transformations that facilitate normative comparisons. Also, this chapter describes how these methods have been applied to data that were generously donated by the ANDI consortium, to form the ANDI database. A description of the contents of the ANDI database is also given.

In the third chapter, multivariate normative comparisons are described, and are extended to include demographic corrections. Also, it is explained how an aggregate database like the one described in chapter two can be used for normative comparisons. An aggregate database is different from standard normative datasets in that there may be differences between contributing studies in how participants perform. In this chapter, a multivariate multilevel regression model is introduced that resolves this issue. A second advantage of this model is that it can be fitted even when many data are missing. Missing data are very common in aggregate data, as some test variables may be completely absent from a particular study. In a simulation study, the appropriateness of the multivariate multilevel regression method is demonstrated. With this method, multivariate normative comparisons with demographic corrections can be made for the most common tests. Another issue related to missing data in aggregate databases, i.e. missing overlap, is left unsolved in this chapter. This issue is addressed in the next chapter.

In the fourth chapter, the method of the previous chapter is extended to solve the issue of missing overlap. If there are two tests that have not been administered together in any of the studies, there is no overlap between the two variables, and it becomes difficult to estimate a multivariate model. This situation would arise with tests that are less commonly administered, as it is more likely that these tests have not been administered together in any of the studies that are included in the aggregate database. Therefore, this prevents the inclusion of less commonly administered tests in the multivariate normative comparison. Two solutions are tested in this chapter, using either a multiple imputation or a factor model approach. This chapter ends with the recommendation that the problem of missing overlap can
best be resolved using a factor model, but only if the factor model is an appropriate description for the included tests. Therefore, the goal of the next chapter is to find an appropriate factor model.

In the fifth chapter, different factor models for neuropsychological tests are compared. These models have been formulated in the literature, and make different distinctions in which test variables measure the same cognitive function. Some models are complex and contain many different cognitive functions, while others are simpler. In this chapter, a factor meta-analysis (Cheung & Chan, 2005) is performed. In this analysis, factor models are fitted to a correlation matrix that is pooled across multiple studies conducted worldwide. From this analysis, a single best fitting model is identified. Next, factor models are fitted to data from the ANDI database. Again, the best fitting factor model is identified. Together with the method described in the fourth chapter, this factor model allows for multivariate normative comparisons with more tests than was possible with the method from the third chapter.

In the sixth chapter, multivariate normative comparisons using ANDI are applied to address a clinical research question. The goal of this study is to classify patients with Parkinson’s disease as either cognitively impaired or not, since impairment at an early stage of the disease is known to predict later development of Parkinson’s disease dementia. With follow-up data that were gathered after three and five years, the performance in the prediction of dementia of the normative comparison procedure described in this thesis is compared to the performance of previously used methods. This thus provides an empirical test of the methods developed in this thesis.

In the seventh chapter, univariate normative comparisons using an aggregate database are discussed. As mentioned before, univariate comparisons can lead to incorrect classifications of cognitive impairment when many different comparisons are performed for different test variables. Therefore, if there is a scenario in which individual test scores are of interest rather than profiles of scores, there needs to be some kind of correction for false positives. In this chapter, several corrections that are described in the literature are discussed, and it is shown how they might be applied with an aggregate normative database. A new method is developed especially for this purpose.

In the eighth chapter, results of the previous chapters are summarized, and limitations and potential solutions are discussed. Possible extensions of the methods and the database, and possible applications outside the current scope of the ANDI project are discussed. The thesis ends with a consideration of how the ANDI project relates to recent developments in psychology.