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

*Citation for published version (APA):*
The goal of this thesis was to improve the reliability of neuropsychological assessment, specifically by improving the normative comparison procedure. The first goal was to provide multivariate normative comparisons, which test the patient’s whole profile of scores. The second goal was to provide normative comparisons that are corrected for age, sex, and level of education. These goals had two requirements. First, a normative database had to be established with many, demographically diverse, healthy participants. Second, a statistical framework had to be developed that allows for demographically corrected multivariate normative comparisons with this new normative database. The statistical framework was the focus of this thesis.

In chapter two, we described how an aggregate normative database can be constructed by combining the data from healthy people from multiple studies. These people may have participated as a control group in a clinical study, or may have participated in a large community study. By combining many such groups of people, data from many different neuropsychological tests can be gathered. All procedures were standardized across tests. This involved two procedures for data cleaning. First, values were discarded that were outside a pre-defined range of allowable scores, which was set beforehand on the basis of clinical expertise. Second, values were discarded that were highly unlikely given participants’ age, sex, and level of education. To select which demographic variables to use in demographic corrections, the Akaike Information Criterion was used. To be able to use parametric statistics, such as parametric normative comparisons, corrected scores would ideally be normally distributed, or transformed to be normally distributed. To select a power transformation to achieve normality, the Box-Cox procedure was used (Box & Cox, 1964). Last, the contents of the ANDI database were described in this chapter.

In chapter three, we described how multivariate normative comparisons can be made using an aggregate database. This required a model that consisted of three parts. First, to include demographic corrections for age, sex, and level of education, a regression model was required to estimate the regression coefficients for these three demographic variables. Second, there may be differences between studies in the scores that healthy participants obtain, for example due to differences in sample selection or test administration between
studies. Therefore, a multilevel model was required, to model these differences between studies. Third, multivariate normative comparisons take into account the relations between scores on different tests. Therefore, the covariance between scores needed to be estimated, and a multivariate model was required. To combine these parts, a multivariate multilevel regression model was formulated. This multivariate multilevel regression has an added advantage, in that it can be fitted with missing data in the test variables. Because of the nature of an aggregate database, large amounts of missing data are to be expected, as tests that were not administered in a particular study have missing values for all participants in that study. With this model, all the components that are required in the multivariate normative comparisons can be estimated: demographically corrected means, variances, and covariances. In a simulation study, performance of the multivariate normative comparisons procedure was evaluated with varying amounts of missing data and between study variance. It was shown that although the model can be fitted using missing data, it cannot if there is missing overlap between tests. This issue was addressed in chapter four.

In chapter four, we described how the model from chapter three can be extended to accommodate missing overlap between tests. There is missing overlap between two tests, if the combination of these two tests has not been administered in any of the studies that are included in the database. This makes the covariance between these two tests impossible to estimate directly. In this chapter, two methods that can solve this problem are identified. The first is multiple imputation, where values are imputed for every missing value. From these imputed values, the covariance can be estimated in a straightforward manner. The second is a factor model approach, where a model for the covariance structure is estimated. This model assumes that the covariance between tests can be described by the dependence of these tests on the same latent variable. In a simulation study, the two methods are compared. The multiple imputation approach keeps the number of false positives under control, but due to underestimation of the covariance between tests, it is less sensitive in detecting impairment than the factor model approach. A precondition for the factor model approach is the appropriateness of the factor model for the data. If the factor model is not appropriate, the number of false positives increases. Therefore, a factor model for neuropsychological tests needs to be established before this model can be applied. This issue was addressed in chapter five.

In chapter five, the fit of different factor models for neuropsychological tests was compared in two studies. In the first study, a meta-analysis, correlation matrices for neuropsychological tests were requested from published studies. The correlation of test scores with demographic variables was partialed out from the correlation between
tests. Subsequently, the correlation matrices were pooled into a single correlation matrix, to which factor models could be fitted. In the second study, factor models were fitted to demographically corrected data from the ANDI database. In both studies, model comparisons showed that the Cattell-Horn-Carroll model as modified by Jewsbury et al. (2016) fitted best. This model was originally developed in intelligence research, and divides cognitive functioning as measured by neuropsychological tests in domains of "Acquired knowledge or crystallized ability", "Processing speed", "Long-term memory encoding and retrieval", "Working memory", and "Word fluency". This is in contrast to other models that divide cognitive functioning in domains of "Attention", "Executive functioning", and "Memory". Because the Cattell-Horn-Carroll model seems to fit data from healthy people well, this model can be used in ANDI to apply the methods developed in chapter four.

In chapter six, the methods developed in this thesis were put to an empirical test. Specifically, the ANDI database and multivariate normative comparisons were used in a re-analysis of longitudinal data from a study on Parkinson’s disease and Parkinson’s disease dementia (Broeders et al., 2013). These data had been analyzed before using conventional (univariate) criteria for Mild Cognitive Impairment in Parkinson’s disease (PD-MCI; Litvan et al., 2012). The goal of the previous study had been to see whether those who fit the PD-MCI criteria at the first measurement occasion would progress to dementia at a later measurement occasion. In this chapter, the results from this study were compared to results obtained with the ANDI database. First, using the univariate PD-MCI criteria with the ANDI database showed more cautious results than the earlier study: Fewer patients were classified as cognitively impaired. This was the case for both patients that later did, and did not develop dementia. Second, application of the ANDI database with multivariate normative comparisons was shown to provide better predictions than using the conventional PD-MCI criteria: They were both more sensitive and specific in predicting who would develop dementia. This provides evidence that the methods described here are indeed useful in improving neuropsychological assessment.

In chapter seven, we returned to the issue of using univariate normative comparisons in clinical neuropsychology. If no correction is used, and many univariate normative comparisons are performed for many different test variables, the number of times that cognition is judged to be impaired in healthy people is increased, a so-called increased familywise error rate. This may have contributed to the lower specificity for the PD-MCI criteria in chapter six. To correct for this increased familywise error rate, correction methods have been developed. A correction method that is frequently used in science, but not so much in clinical practice, is the Bonferroni correction. This correc-
tion decreases false positives, but can hurt the chances of detecting impairments in those who are truly impaired. In this study, more sophisticated correction methods were discussed and compared in a simulation study, specifically for the situation where patients are compared to an aggregate database. A new stepwise method performed better than the Bonferroni correction in detecting impairments in many settings, but did show an increase in false positives if many data were missing. Therefore, it is too early to fully endorse either method.

This thesis was accompanied by the establishment of the ANDI database and website. In this project, 84 generous contributions from research groups across the Netherlands and Belgium yielded data from 27,000 participants. In the ANDI project, the methods described in chapter two and three have been implemented. The website will be extended using the method from chapter four, with the model from chapter five.

8.2 POTENTIAL IMPROVEMENTS

The models in this dissertation were focused on multivariate distributions, and multilevel and factor model approaches. These approaches have had a number of advantages, in terms of the estimability of parameters in the light of many missing data points and differences between studies. However, these approaches brought with them a number of assumptions, in terms of linearity, equality of variances across different levels of demographic variables, and normality of the data. Voncken, Albers, & Timmerman (2017) proposed a powerful method for norming data that does not make these assumptions. However, their method is not multivariate, is not yet tested for very high percentages of missing values, and has not yet been developed for aggregate databases. One future direction could be to borrow the best from both methods, to arrive at multivariate normative comparisons while relaxing some of the assumptions where necessary.

The multivariate normative comparison procedure provides a single dichotomization into impaired and not impaired for a whole profile of test scores. This information is relevant in many clinical and research settings. However, in other settings, more detailed information on the nature of the deviation is needed, or a measure of the severity of impairment is needed. An option would be to study each of the test scores separately using univariate normative comparisons. Therefore, one approach would be to further improve the univariate approach from chapter seven, in order to make sure that it keeps the number of false positives low with missing normative data. Another approach would be to make multiple multivariate normative comparisons for parts of the profile, for example only comparing tests on two domains multivariately, ignoring the remainder of scores, and then comparing tests on two other domains multivariately. Whether
this increases sensitivity to certain impairments, and how to control for false positives in this scenario are topics for future research.

The question remains how rare a patient’s score or score profile has to be in healthy people before the patient’s cognitive functioning can be classified as impaired. In this thesis, common thresholds were used, such as using the criterion that if a score this low is obtained by 5% of healthy people or less, we consider cognitive functioning to be impaired. How this threshold is set determines both sensitivity to real cognitive impairments, and the chance of finding a false positive. Therefore, it is important to use a threshold that maximizes performance in both respects. This 5% is therefore not set in stone, and should be considered a starting point. One reason that the best possible threshold has not been determined before, is that it is highly dependent on the context of the assessment. In some contexts, the base rate, i.e., the number of people with real cognitive impairments, will be higher than in other contexts. In a context with a high base rate, say at the intake of a memory clinic where patients with subjective complaints are invited, 5% may be too strict a threshold for impairment, and may result in many cognitive impairments being missed. In a context with a low base rate, say a screening of patients who have fallen recently, 5% may be too lenient a threshold for impairment, and may result in many false positives. Therefore, a fruitful extension of the present thesis would be to extend the multivariate normative comparisons method using Bayes’ rule (Gavett, 2015) to take into account differences in base rates between different contexts.

8.3 Extensions of the Method

This thesis has focused on cross-sectional data from healthy people that completed a test battery for the first time. This makes this type of database useful to clinical neuropsychologists interested in setting diagnoses, and characterizing deficits in patients who complete a test battery for the first time. However, in for example treatment settings, clinical neuropsychologists also evaluate a patient’s test scores at multiple occasions. To evaluate whether a patient’s progression over time is different than observed in healthy people, an aggregate database of longitudinal data would ideally be built as well. There are multiple ways to envision such a longitudinal database.

One option would be to focus on a single retest session, for example after three months, for which normative data could be collected. Then, the decrease or increase of scores that patients show from baseline to this three-month follow-up can be compared to that of healthy people. This option could be difficult to implement if there are few studies that have retested healthy people after three months. Also, the application of this database would be limited to patients who are retested after three months. An advantage would be that the statis-
tical framework of normative comparisons in this setting is already available in the form of the Reliable Change Index (Jacobson & Truax, 1991), which could also be extended to a multivariate version.

A second option would be to collect data from healthy people who were tested at varying time intervals. This would allow for the inclusion of more datasets, and could be more widely applicable because patients are tested at varying time intervals as well. However, the statistics for the modeling of changes over time would be more involved. From the longitudinal data, progressions of scores over time, i.e., the slope of the regression line, could be estimated for every healthy person. The patient’s slope could then be compared to healthy participants’ slopes in the by now familiar normative comparison procedures. Slopes of multiple test variables could then be analyzed using multivariate normative comparisons in the same way. One issue would be that time is not the only factor influencing differences between scores between measurement occasions, as practice effects from the measurement itself may improve scores over time (McCaffrey & Westervelt, 1995). This makes it difficult to pool data from a study with a measurement after three months, and a study with measurements every week.

Apart from longitudinal data, data from samples other than healthy people could be useful to clinical practice. This thesis was fundamentally focused on normative comparisons, that is, comparing a patient’s scores to scores obtained by healthy people. An alternative would be to compare a patient’s scores not only to scores obtained by healthy people, but also to scores obtained by clinical groups. With such data, a different statistical approach would be needed, to classify whether a patient’s cognitive functioning is more similar to that of healthy people, or more similar to that of a particular clinical group. This would require that large samples of participants are available from different clinical groups, to best be able to make this distinction. We can be sure that studies are available amenable to the goal of pooling clinical groups, as in clinical studies often extensive test batteries are administered. Therefore, it would be worthwhile to add a database of patients from different studies to the existing database of healthy participants.

8.4 Potential applications

The focus of the project was on establishing a Dutch and Belgian normative database of tests used in clinical neuropsychology to diagnose cognitive impairment in adults. However, every step is generally applicable to normative comparisons in other domains. This is true for the pre-processing steps in chapter two, and is also true for the models of chapters three and four. It should be noted that the code required to perform these methods is available online, as is the code.
to run the website on which the normative comparisons can be made. Because of the generality of the methods proposed here, there are many applications that could follow naturally from this thesis.

First, the methods are not specific to Dutch and Belgian data. Therefore, this project can be replicated in other countries and/or regions. The availability of clinical neuropsychology norm data was already quite good in the Netherlands and Belgium, as test publishers provide high quality norms for this language area. Therefore, an aggregate database of normative data would be even more valuable in countries where there are fewer high quality norms available. One suggestion would be to keep the regions small for a single database: If studies from too many countries are aggregated, the variance between studies becomes large due to differences between countries in language, culture, and test versions.

Second, the methods are not specific to adults. Therefore, this project can be replicated for developmental clinical neuropsychology data as well. In fact, several donations to the ANDI database have already been made for children’s data. It is not advisable to create a single database for children and adults, as the tests that are typically used are different between children and adults. Adding data from these age groups together would create a problem of missing overlap that is more severe than what was discussed in this thesis. Therefore, a new aggregate database of normative data for children would be advised.

Third, the methods developed here are not limited to applications in clinical neuropsychology. Clinical neuropsychology has the advantage that tests are highly standardized in the way they are administered, and in the way the outcomes of these tests are scored. This enables the combination of data across different sources. If there would be more flexibility in the administration procedure, this would contribute to more variance between studies. However, there are many fields within and outside of psychology where standardized tests and standardized outcomes are common, for example in clinical psychology (Clark & Watson, 1995), personnel psychology (Bartram, 2008), and educational science (Delandshere, 2001). For each of these fields, a normative database of test scores could be established, so multivariate normative comparisons can be used to classify whether a single case is typical, or different from the norm.

Fourth, although so far the discussion has only been about test scores, normative comparisons can be extended outside the domain of testing. Biomarkers would be one domain where an aggregate normative database would be useful. With such a normative database, for example a patient’s blood pressure and heart rate variability could be compared to those of healthy people (Morrison & Morris, 1959; Tsuji et al., 1996). This could also be done in a multivariate normative comparison, taking the correlation between blood pressure and heart rate
variability into account. Other biomarkers that would be amenable to normative comparisons could be brain indices (Dubois & Adolphs, 2016), obtained using fMRI, EEG or PET. Brain indices produce large amounts of data from different locations in the brain, which increases the chance of a false positive if every location is considered separately. Therefore, multivariate normative comparisons would be useful in this field as well.

8.5 Themes

With this thesis and the broader ANDI project, we wanted to show that an aggregate normative database can provide clinical neuropsychologists with the data they need to apply statistically advanced techniques, which can improve diagnostics in clinical neuropsychology practice. Three themes that pervade this thesis are treated next.

The first theme is data sharing. The establishment of aggregate normative databases is only possible if there is broad willingness in the research community to share data. The ANDI project is very fortunate that there is a culture of cooperation in the clinical neuropsychology community in the Netherlands and Belgium. We hope that this will also be the case for future data aggregation ventures.

The second theme is integration of substantive and methodological fields within psychology. It generally takes a long time before newly developed statistical methods become available to other researchers, and before those methods that are available in research become available in clinical practice. With the advent of the free R statistics software with its productive community of developers, new statistical tools become freely available every day. However, the data and the know-how to apply these newly developed tools are also necessary. With the ANDI project, we hope to have crossed the divide between methodological development and substantive questions by developing a user-friendly website with which clinicians and researchers can analyze their own patient data.

The third and last theme is valorization. In recent years, there has been a call for science to become more useful to society at large. Scientists are asked to come up with studies that result in products that can be used, thereby providing value outside science. One criticism of this call for so-called valorization is that it could detract from research for which it is not immediately clear what the value is to society, but which may be valuable in its own right, or which may prove valuable in the long run. In the ANDI project, the data were not collected with valorization specifically in mind. However, they were re-used to create a product that is immediately useful to clinicians and patients. This seems like an ideal example of valorization.
8.6 conclusion

Neuropsychological assessment is an important part of clinical care and clinical research. When normative comparisons are reliable, we can use neuropsychological assessments to discover impairments in a patient’s cognitive functioning, and to discover cognitive benefits and side effects of new treatments. Therefore, it is important to make sure that normative comparisons are as reliable as possible. In this thesis and the ANDI project, we improved normative comparisons, by developing a normative database and by developing a statistical framework to make normative comparisons with this database. These developments made multivariate normative comparisons, and more accurate demographic corrections, available to clinical neuropsychologists in practice and research.