TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY
Julkaisu 683 • Publication 683

Esa Wallius

Increasing Statistical Power in Brain PET Studies by Image Analysis Methods: Applications for Drug Development

Tampere 2007

Tampereen teknillinen yliopisto. Julkaisu 683
Tampere University of Technology. Publication 683

Esa Wallius

Increasing Statistical Power in Brain PET Studies by Image Analysis Methods: Applications for Drug Development

Thesis for the degree of Doctor of Philosophy to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 12th of October 2007, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2007

ISBN 978-952-15-1853-9 (printed)
ISBN 978-952-15-1880-5 (PDF)
ISSN 1459-2045

Abstract

This thesis provides means to increase the statistical power of analysis in brain positron emission tomography (PET) neuroreceptor studies. Despite its importance, this topic has remained relatively unexplored. The most obvious way to enhance the power is to increase the sample size, i.e. the number of observations. However, with PET, there are both financial (high cost of a study) and ethical reasons (radiation dose to the patient) that restrict substantial increase in the sample size.

We searched for ways to increase the statistical power in the analysis of PET brain studies by decreasing the measurement error (specifically, variation due to the measurement process) as well as by using a sensitive statistical technique of analysis. Our procedure was as follows: firstly, we reduced probably the greatest source of variation – the manual steps – by introducing an entirely automatic method for region-of-interest delineation. Secondly, we performed phantom simulations to determine the combination of image reconstruction methods and physiological model calculation methods that produced the least variation (and bias). Thirdly, we applied a statistical method which produces satisfactory statistical power with a very limited sample size. In the first two methods, we showed the enhancement in power by the better reproducibility and reliability, as well as lower variability, of the methods we applied. In the third method, the gain in power was illustrated by comparing the results with the standard method of analysis. If all three methods are performed, it may be possible to get satisfactory statistical power even with a limited sample size and possibly restrict the sample size even more.

Our application was in PET receptor occupancy studies, which have been found to be useful in guiding the dose-finding procedures in early drug development. In addition to the gain in power to detect the effects of a drug, cost savings and an increase in speed can be achieved with the methods introduced in this thesis. The cost savings relate to pre-planning: realistic simulations with numerical dynamic phantoms may save many

expensive pilot PET studies on humans. The increase in speed was obtained by automation of the region-of-interest delineation.

Preface

This thesis was carried out during 2001-2007 in the Institute of Signal Processing of Tampere University of Technology, in close collaboration with the Turku PET Centre, University of Turku/Turku University Central Hospital.

I owe my deepest gratitude to my supervisor, Professor Ulla Ruotsalainen, for her inspiring and encouraging guidance, as well as for sharing her vast knowledge of medical image analysis. I am indebted to my instructor, senior researcher Jussi Tohka, PhD, the co-author of three publications of this thesis, for numerous invaluable discussions on methodological issues and for excellent advice.

The pre-examiners of this thesis, Professor Kenneth Nordström, University of Oulu and Balázs Gulyás, MD, PhD, Karolinska Institute, Stockholm deserve sincere thanks for their careful reading of and constructive comments on this thesis.

I thank our collaborators in the Turku PET Centre, Professor Jarmo Hietala, Jussi Hirvonen, MD, PhD, Sargo Aalto, MSc, Vesa Oikonen, MSc, and Mikko Nyman, MD, for fruitful co-operation in preparing the joint publications and for providing PET data.

I also wish to thank Anu Juslin, MSc, Sakari Alenius, PhD, Johanna Eskola, MSc, Margarita Magadán Méndez, MSc, Antti Happonen, PhD, Evgeny Krestyannikov, MSc, Jyrki Möttönen, PhD, Sari Peltonen, PhD, and Jouni Mykkänen, PhD, as well as all the present and past M^2OBSI group members for assistance whenever I needed it. My parents, my sister and my good friends deserve thanks for their encouragement.

Lastly, I would like to thank the funding organizations, the Tampere Graduate School of Information Sciences and Engineering (TISE), the TEKES Drug 2000 Technology Programme and the Academy of Finland, Finnish Centre of Excellence Programme (2006 - 2011).

Tampere, 17th of September, 2007 Esa Wallius

Supervisor: Professor Ulla Ruotsalainen
Institute of Signal Processing
Tampere University of Technology

Instructor: Senior Researcher Jussi Tohka, PhD
Institute of Signal Processing
Tampere University of Technology

Pre-examiners: Professor Kenneth Nordström
Department of Mathematical Sciences / Statistics
University of Oulu

Balázs Gulyás, MD, PhD
Psychiatry Section
Department of Clinical Neuroscience
Karolinska Institute, Stockholm

Opponents: Professor Lauri Tarkkonen
Department of Mathematics and Statistics
University of Helsinki

Zsolt Cselényi, PhD
R & D, AstraZeneca AB, Södertälje and
Psychiatry Section, Department of Clinical Neuroscience
Karolinska Institute, Stockholm

Tampere University of Technology
Department of Information Technology
Institute of Signal Processing
Methods and Models for Biological Signals and Images (M^2OBSI) Group
Tampere Graduate School in Information Science and Engineering (TISE)

Contents

List of publications

This thesis consists of an overview and the following five publications. The publications are referred to in the text as (Publication *), where * denotes a roman numeral.

Publication I Wallius E, Nyman M, Oikonen V, Hietala J and Ruotsalainen U. Voxel-based NK1 receptor occupancy measurements with [^{18}F]SPA-RQ and positron emission tomography: A procedure for assessing errors from image reconstruction and physiological modeling. Molecular Imaging and Biology, 9(5):284-294, 2007.

Publication II Tohka J, Wallius E, Hirvonen J, Hietala J and Ruotsalainen U. Automatic extraction of caudate and putamen in [^{11}C]raclopride PET using deformable surface models and normalized cuts. IEEE Transactions on Nuclear Science, 53(1):220-227, 2006.

Publication III Wallius E, Tohka J, Hirvonen J, Hietala J and Ruotsalainen U. Evaluation of the automatic three-dimensional delineation of caudate and putamen for PET receptor occupancy studies. To appear in Nuclear Medicine Communications, 2007.

Publication IV Wallius E, Tohka J, Hirvonen J, Hietala J and Ruotsalainen U. A method for automatic extraction of striatal structures for PET dose-finding studies. In proc. of IEEE Medical Imaging Conference (MIC2006), pp. 3189-3194, 2006.

Publication V Aalto S, Wallius E, Näätänen P, Hiltunen J, Metsähonkala L, Sipilä H and Karlsson H. Regression analysis utilizing subjective evaluation of emotional experience in PET studies on emotions. Brain Research Protocols, 15:142-154, 2005.

List of abbreviations and mathematical notations

~ distributed as

BP binding potential

CV coefficient of variation

DSM-IS dual surface minimization – inner surface

DSM-OS dual surface minimization – outer surface

E(.) expected value of (.)

FBP filtered back-projection reconstruction

FDR false discovery rate

FWE familywise error rate

GLM general linear model

ICC intraclass correlation coefficient

MR magnetic resonance

MRP median root prior reconstruction

n number of observations, i.e. sample size

NAD normalized absolute difference

P(.) probability of (.)

PET positron emission tomography

ROI region of interest

S sample standard deviation

SEM standard error of measurements

SPM statistical parametric mapping

SRTM BF simplified reference tissue model with basis functions

SVC small volume correction

TRV test-retest variability

Var(.) variance of (.)

$\bar{X}$ sample mean

Chapter 1 Introduction

Whenever measurements are made, errors always occur due to the measurement process. These errors are taken into account in the statistical analysis, which often deals with dividing the variation in variables of interest (i.e. dependent or response variables; treated as random variables) within the framework of the general linear model (GLM; e.g. Rencher, 2000; Cox and Hinkley, 1974). This division is performed between the experimental factors (i.e. independent variables), and the random errors that are often assumed to be independently and normally distributed with zero mean. In scientific experiments, the errors can be characterized as measurement errors that divide into random error (i.e. variability, or more commonly, noise) and systematic error (bias). Random error can further be classified into natural biological variation and variation due to the measurement process, i.e. methodological variation (e.g. Snedecor and Cochran, 1967, p. 164-165; Carroll et al., 1995, p. 2). In many cases, statistical analyses involve a hypothesis testing framework (Neyman and Pearson, 1933), but there are multiple types of analyses that do not necessarily; e.g. principal component analysis and statistical shape analysis. In hypothesis testing, a null hypothesis and an alternative hypothesis are formulated. The latter is usually the hypothesis for which the researcher is searching support. The statistical testing is often based on specific distributions (Fig. 1). The statistical power means the probability that the test will reject a false null hypothesis. In other words, it gives the probability that truly existing differences will be detected. Assuming the correct model, there are three issues affecting the statistical power: the significance criterion (α), the reliability of the sample results, including its most important factor, the sample size, and the effect size, i.e. the degree to which the phenomenon exists (Cohen, 1969). In other words, this can be formulated as how much of a difference the experimental factors make.

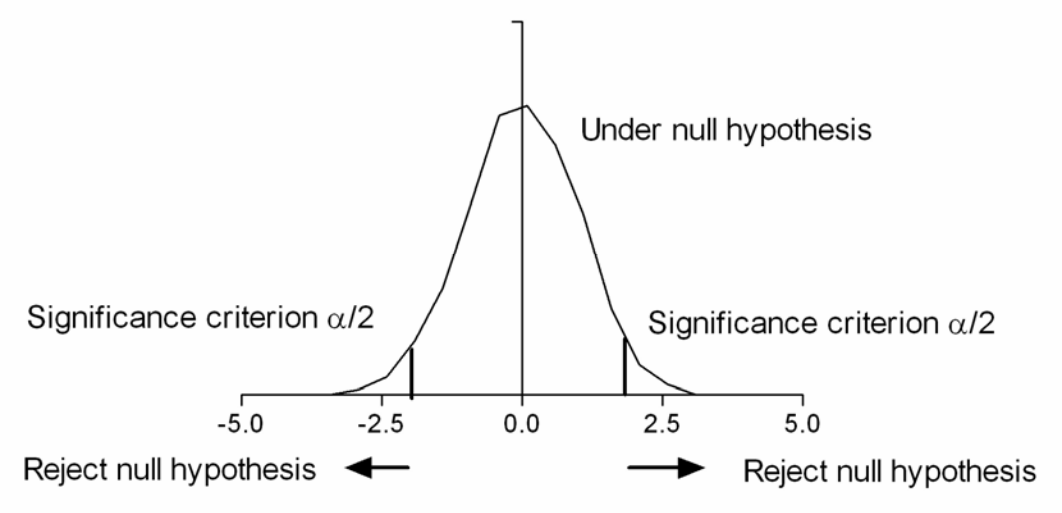

Figure 1. An example of statistical hypothesis testing. Histogram of a sample of 3000 observations from standard normal distribution (mean 0, variance 1) under null hypothesis is depicted. Values more extreme than the significance criterion (α/2) show when there is enough evidence to reject the null hypothesis.

The motivation for this thesis was to attempt to increase the statistical power of the analysis of functional positron emission tomography (PET) imaging data. PET is increasingly used in drug development, in phase one of three-phase clinical trials. As a specific application, we dealt with receptor occupancy studies (Laruelle, 2000; Passchier et al., 2002) that guide the dose-finding procedures in early drug development. Increasing the power is relevant as the PET data are very noisy; see Pajevic et al. (1998) for noise characteristics of PET images. In this work, we did not try to enhance the statistical power by increasing the sample size, which is the most obvious choice to improve the reliability of sample results. The reason that we did not do this is related to the limitations of PET imaging that are described below. On the contrary, we aimed at reducing the errors derived from the applied methods, and also increased the statistical power by efficient techniques of analysis. In brief, we used signal and image analysis methods to improve the statistical analysis. Power analysis (e.g. calculation of sample size with fixed power) was not performed as the actual statistical analysis did not, in general, play a

remarkable role in the publications of this thesis. The improvement in power was shown by other indicators, such as reduced variability, improved reliability, or better reproducibility.

Chapter 2 The statistical power and related concepts

2.1. Statistical inductive inference

Statistics has, in some circumstances, been called a combination of mathematics, philosophy and computer science. The interpretation of the results of the statistical tests is based on inductive inference, which is a branch of logic. The formal theory of inductive inference was developed by the computer scientist, Solomonoff (1964a; 1964b). In inductive inference, the premises, or evidence statements, can be true, whereas the inferred conclusion may be false without logical contradiction: the conclusion is "evidence transcending" (Mayo and Cox, 2006). This is because probability plays a role in such inferences. Statistical inference (reasoning) may be regarded as a statistical version of the valid form of the argument called in deductive logic, *modus tollens*, i.e. proof by contra-positive (Mayo and Cox, 2006). In this form of inference, the denial of a hypothesis H is inferred as follows: If H then E. E is false. Therefore H is false. In statistical inductive inference, it is argued from the particular to the general, more specifically from a sample to a population.

2.2. General issues on statistical hypothesis testing

Statistical inference often involves hypothesis testing. In this thesis, we will treat hypothesis testing in the frequentist inference framework (e.g. by Neyman-Pearson and Fisher; see specific examples below). Frequentist statistics is based on long-run frequencies, one of the most difficult conceptual problems in statistical inference (Hacking, 1965). Hypothesis testing (or model selection) can, however, also be applied in other statistical and information theoretic frameworks, such as Bayesian inference (Aitkin et al., 2005; Lee, 1997), the minimum description length principle (Rissanen, 1978; Barron et al., 1998) and the Bayesian-like minimum message length approach (Wallace and Boulton, 1968; Wallace and Boulton, 1987). In this thesis, we deal with hypothesis testing in a test-retest (two repeated measurements from each subject) situation to

examine and reduce within subject and between subjects variability. The ultimate goal is to enhance the statistical power by reducing variability.

2.2.1. Hypothesis testing approaches by Neyman-Pearson and Fisher and their differences

The approach by Neyman-Pearson (Neyman and Pearson, 1933) is usually called just "hypothesis testing". On the contrary, Fisher's approach (Fisher, 1934; Fisher, 1973) is commonly titled "significance testing", although Fisher was not the originator of significance testing (Royall, 1997). A central term in significance testing is the notion of p value. It is not a typical probability about the null hypothesis, but shows the probability that we have observed the data in hand, or more extreme data, given that the null hypothesis is true (Fisher, 1934; Cohen, 1994). The practical statistical analysis is currently typically a blend of both of these approaches, also in functional neuroimaging (e.g. Turkheimer, 2004). However, these different approaches may possibly be united (Lehmann, 1993).

In this thesis, the most important difference between the two approaches concerns the statistical power. A central consideration of the Neyman-Pearson theory is that in addition to the null hypothesis, also the exact alternatives against which it is to be tested must be specified. In terms of alternatives, one can define the statistical power of the test. This approach is important in assessing the chance of detection of an effect, i.e. the departure from the null hypothesis. However, Fisher, although he acknowledged the importance of power, denied the possibility to assess it quantitatively. In his opinion, this was due to the unknown alternative. Other contradictions between the Fisher and Neyman-Pearson approaches were, e.g. conditioning, the use of likelihood (as opposed to probability), continuous versus fixed p values (significance criterions), decision-making, and philosophical controversies concerning just how "inductive" the statistical inference actually is. Especially the philosophical differences between the two schools were – and even still are – debated and argued extensively. All these issues are widely discussed in Lehmann (1993).

2.3. Hypotheses, errors, and analogues in statistical testing

2.3.1. Null hypothesis and alternative hypothesis

The statistical hypothesis test begins with the definition of the null hypothesis (H_0) and the alternative hypothesis (H_1), which is of greatest interest. This alternative hypothesis is not proved (and never can be), but the null hypothesis is tested instead. Usually it is of interest to find a significant departure from the null hypothesis to "reject" it. However, it should be kept in mind that the null hypothesis can not be proved correct either, although the data show no evidence for rejecting it. Let us take a simple example: testing the difference variable of two consecutive measurements from the same subjects measured at different time points (i.e. a test-retest study) using the paired t test. The null hypothesis states that the difference μ between the measurements at population level is some predefined value μ_0, whereas the alternative hypothesis argues that the difference is not μ_0, i.e.

$$\mathrm{H}_0 : \mu = \mu_0,\ H_1 : \mu \neq \mu_0. \tag{1}$$

The test statistic is

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1), \tag{2}$$

where t(n-1) is t distribution with n-1 degrees of freedom. A formal definition of degrees of freedom is D(K) – D(H) (Good, 1973), where D(K) is the dimensionality of a broader hypothesis and D(H) is the dimensionality of the null hypothesis. From this example, we can also infer that we are dealing with a two-sided test; a one-sided test would only consider deviation in one direction, not both.

2.3.2. Type I error, type II error and the statistical power

Statistical hypothesis testing is often based on specific probability distributions, depending on the question (in the case of the previous paragraph the t distribution). A simple hypothesis completely determines the probability distribution, whereas a composite hypothesis asserts only that the distribution belongs to a specified set of distributions. In this testing framework, one can make two kinds of errors: type I error (α) and type II error (β). A type I error means that a true hypothesis is rejected or, in a milder sense, the risk of a false decision. This can be formulated as P("reject H_0"| H_0) = α. Usually the type I error is adequately taken into account in the analysis by fixing it to some significance criterion (conventionally 0.05, but this is arbitrary). A Type II error occurs when a false hypothesis is accepted and is defined as P("accept H_0"| H_1) = β. The type II error is, unfortunately, not typically adequately considered in the statistical analysis. This occurs especially when the sample size is small, and no power calculation for the required sample size has been carried out. Then, the type II error may be considerable, but this can be avoided if prior power calculations have been carried out. The statistical power is then defined as 1-β, and it indicates the probability that the test will reject a false null hypothesis. In power calculations, the statistical power is commonly fixed to 0.8, thus β is 0.2. However, other values are also used, depending on the relative severity of type I and type II errors in the specific application. It is known that there is a trade-off between type I error and type II error. Most of the issues in this paragraph are covered in Lindgren (1968).

2.3.3. Analogous examples of hypothesis testing in other fields

There is an analogy between the statistical hypothesis testing and, e.g., with juridical issues, doping control in sports, and diagnostic tests in medicine. The famous statement in law "Innocent unless convicted" is analogous to a null hypothesis. Contrary to statistical testing, in law attempts are made to make the type I error negligible, i.e. approach zero. This tendency not to sentence innocent people inevitably leads to inflated type II errors, i.e. many criminals have not been sentenced. A similar strategy concerns

doping testing in sports. The inflated type II error makes it, e.g. impossible to know definitely which records are actually made with the aid of performance-enhancing drugs. On the contrary, concerning diagnostic tests in medicine, type II errors may be more harmful than type I errors. In statistical signal processing, the problem is related to finding the signal in noisy data, i.e. the null hypothesis is "noise only" (Kay, 1998).

2.3.4. Binary classification tests

The error types in statistical hypothesis testing can be easily described as a binary classification test (Table 1). There are differences in terminology, i.e., the same issue is formulated in different words for statistics, signal processing (Kay, 1998) and diagnostic tests (Pepe, 2003). In this thesis, we will mostly use statistical terms.

		Actual condition		
		True	False	
Test result	Positive	True positive (TP)	False positive (FP; Type I error)	Positive predictive value (PPV) = TP / TP + FP
	Negative	False negative (FN; Type II error)	True negative (TN)	Negative predictive value (NPV) = TN / TN + FN
	Statistics	Statistical power 1-P(Type II error)	1-P(type I error)	
	Signal processing	P(detection)	1-P(false alarm)	
	Diagnostic tests	Sensitivity	Specificity	

Table 1. Binary classification test: result of a test versus actual condition. Also the relation of type I and type II errors to terms in the classification test are described. Distinct terms between statistics and signal processing and diagnostic tests are also presented.

2.4. Factors affecting the statistical power

The power of a statistical test is dependent on three factors (Cohen, 1969; Kraemer and Thiemann, 1987; Stevens, 1986). The first factor is the α level set by the experimenter. This is also called the significance criterion. The second factor is the reliability of the sample results, i.e. the expected accuracy to estimate the appropriate population value. Reliability may or may not be directly dependent on the unit of measurement, the population value, and the shape of the population distribution. However, the reliability is always dependent on the sample size, i.e. the number of observations. This is why the central interest in power calculations lies in the sample size. The third influencing factor is the effect size. This can mean, e.g. how much of a difference the treatments cause, or the extent to which the groups differ in the population (concerning the dependent variables). In other words, it is described as the degree to which the phenomenon exists. For example, in the paired t test, the effect size is written as

$$d = \frac{(\mu_D - \mu_0)}{\sigma}, \tag{3}$$

where d is the effect size, μ_D is the difference in means between the experimental and the control condition, μ_0 is the difference in means under the null hypothesis and σ is the standard deviation for the difference in means in the population (Kramer and Thiemann, 1987). The paired t test is used as an example in this thesis, as it is a very common statistical test in PET neuroreceptor studies (see Aston et al., 2000). The assumptions of the paired t test include the independence and normality of the paired differences (residuals). Potential assumption violations include the lack of independence within a

sample, outliers, interactions between pairs and treatments, differing skewness between the two populations, non-normality and problems with small sample sizes.

2.5. Power calculations, i.e., power analysis

Statistical power, significance criterion, sample size, and effect size are so related that any one of them is a function of the other three (Cohen, 1969). This means that when any three of them are fixed, the fourth is known. This makes formally possible four types of power analysis, of which the sample size calculations to obtain specified power are probably most applied. The results of power calculations depend on the statistical tests and designs (see Cohen, 1969 for power tables in various statistical tests), but we will not go into details here.

The power analysis is commonly performed in the planning stage of a study (Kraemer and Thiemann, 1987). This is the intended and recommended form for power analysis to control the balance of type I and type II errors. However, retrospective power analysis can be performed as well, although it is not universally accepted (Kraemer and Thiemann, 1987). In such cases, the power is typically low, owing to, e.g. small sample size. Power analysis is then only used to show that the absence of a significant effect is probably due to low power.

2.6. Visualizations of types of error and power

2.6.1. Receiver operating characteristic

One way to visualize the relation between type I error and power is to use the receiver operating characteristic (ROC) (Hanley and McNeil, 1982; Pepe, 2003) that is frequently used, e.g. in diagnostic tests especially in radiology, and in detection theory in signal processing (Fig. 2). However, ROC is not widely used in power analysis in statistics. This is probably because, in statistics, only fairly small values of α are usually of interest. On the contrary, ROC analysis covers all the possible values of α and statistical power.

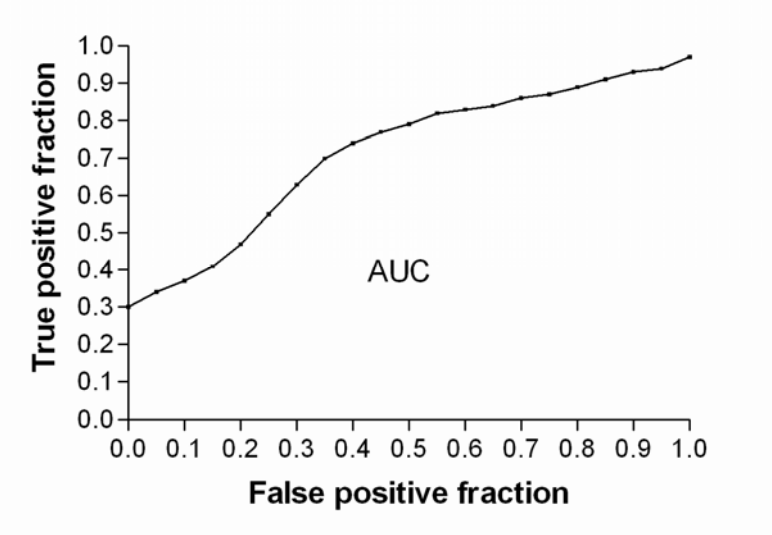

Figure 2. Example of a ROC curve. The statistical power (true positive fraction) is on the vertical axis and type I error (α; false positive fraction) on the horizontal axis. Usually the area under the curve (AUC) is of interest in ROC analysis.

2.6.2. Effect size versus power

In statistics, statistical power is typically plotted against the effect size (e.g. Erdfelder et al., 1996; Fig. 3). Of course, power can also be visualized against the sample size (see Section 3.1; Fig. 4).

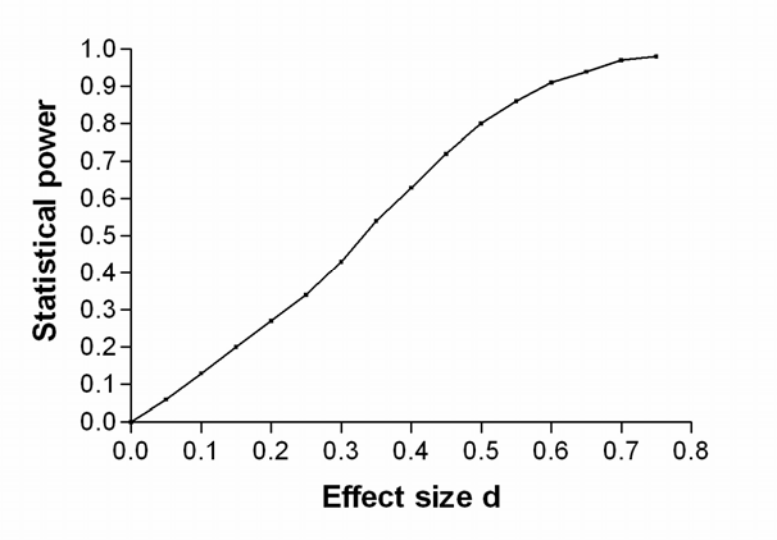

Figure 3. The relation between the effect size (d) and the statistical power. When the effect size increases, the statistical power always increases as well. The values for the figure have been taken from Fig. 5 of Erdfelder et al. (1996). The test is one-sided t test for means, the sample size is 100 and α is 0.05. Effect sizes from 0 to 0.2 have been interpolated by assuming a linear increase in power.

2.7. Correction for multiple tests

Correction for multiple comparisons, or tests, is developed for simultaneous testing of multiple null hypotheses. In these circumstances, the p values have to be corrected to control the familywise error rate (FWE), i.e. to keep the risk of a false decision in a single hypothesis at a predefined level α. FWE is defined as P(N>0), where N is the number of false decisions (Hochberg and Tamhane, 1987; Holmes, 1994). There are also various other error rates, but we do not deal with them here (see Holmes, 1994; Nichols and Hayasaka, 2003). Recently, a fundamentally different approach to control for multiple tests, the false discovery rate (FDR) has been presented (Benjamini and Hochberg, 1995). FDR has also been introduced in the context of neuroimaging (Genovese et al., 2002; see Chapter 5).

2.8. Other considerations

A vast amount of criticism has been presented on statistical hypothesis testing in many application areas (e.g. Turkheimer et al., 2004; Krantz, 1999). At the present time, there is a recommendation by the International Committee of Medical Journal Editors to avoid relying solely on statistical hypothesis testing. A partial solution to this is to use confidence intervals (invented by Neyman) accompanying the p values. Although confidence intervals are connected with p values, they provide valuable information about the spread and magnitude of the effects. If one is not satisfied with these kinds of intervals, either intervals based on the maximum likelihood theory (invented by Fisher) or Bayesian statistics can be applied. It should also be noted that statistical significance is different from practical significance, e.g. clinical significance. In final decisions, practical significance should be more important than statistical significance. For example, if the result is statistically significant, but the confidence interval includes a clinical significance limit, the result is not adequate.

Chapter 3 General means to increase the statistical power

The primary motivation for increasing the statistical power is to obtain more reliable conclusions resulting from statistical tests by means of advance planning. Careful planning is always beneficial in research, and necessary in drug development. Planning a study involves: 1) Specification of research goals in precise and realistic terms, 2) Identification of the design and measurement options available to address the research questions; see Publication I for specific application, 3) Evaluation of the resources, i.e. time, personnel and funding, available to the project. Statistical power considerations are only then used to compare the potential consequences of such alternatives. In addition, power analysis guides the most feasible and cost-effective choice in a particular research setting (Kraemer and Thiemann, 1987).

3.1. Sample size and related issues

The most effective way to detect the intervention effects is to increase the sample size (Fig. 4). As the sample size increases, power also increases, because the standard error of the mean decreases by the square root of the sample size (Lindgren, 1968). The effect of increased sample size on power depends on the statistical design (see, e.g. Cohen, 1969). However, there may be, e.g. financial, ethical or patient availability restrictions in increasing the sample size.

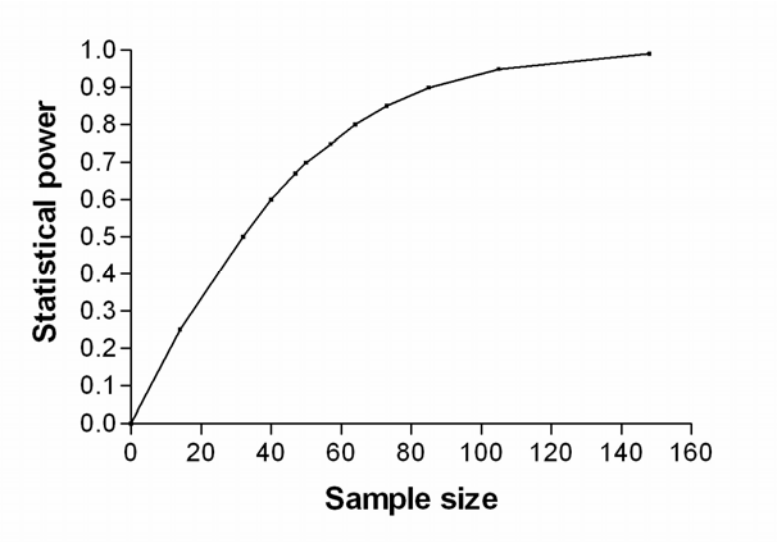

Figure 4. Sample size visualized as a function of statistical power. The increase in the sample size always implies enhanced power to detect the effects, other issues being similar. This example is taken from Cohen (1969, p. 53). The test is a two-group t test, the significance criterion is 0.05, and the effect size is moderate, i.e. 0.5.

Meta-analysis ("multi-trials") can be used to increase the sample size and thus the statistical power of the test (see Hedges et al., 2001) both in exploratory research and controlled, randomized clinical trials. Meta-analysis can be broadly defined as a statistical technique for combining estimated treatment effects from independent comparable studies (DerSimonian and Kacker, 2007). Concerning the statistical model, there are two types of meta-analysis: the fixed effects model (e.g. Mantel and Haenszel, 1959; Peto et al., 1976; Peto et al., 1977) and the random effects model (e.g. DerSimonian and Kacker, 2007). The difference between the two lies in dealing with the treatment effect: the fixed effects model assumes the treatment effects to be the same in different studies, whereas the random effects model treats the variation in treatment effects as a random variable (O'Rourke, 2006). A meta-analysis can be done either from the literature or from individual patient data. Another, related, way to increase the sample size and power is to use randomized multicentre clinical trials ("mega-trials") (Meinert,

1981; Fleiss, 1986). In a multicentre clinical trial, researchers from different locations are brought together to specify the crucial research questions, using the rationale and justification provided by prior single-location randomized clinical trials (Kraemer and Robinson, 2005). An interesting discussion about the pros and cons as well as comparison between meta-analysis and multicentre clinical trials can be found in Shrier et al. (2007). To summarize, it was suggested that multicentre clinical trials have a small advantage with respect to confounding by chance, but meta-analysis may represent a significant advantage in introducing inherent heterogeneity from smaller trials. However, the preference for meta-analysis or multicentre clinical trial is related to whether methodological heterogeneity is considered beneficial or detrimental.

Further improvement in power can be gained by increasing the degrees of freedom for better estimation in a statistical test. In many cases, the degrees of freedom depend on the number of observations and subjects. The increased degrees of freedom will yield better statistical power, i.e. sensitivity (e.g. Friston et al., 1999).

3.2. Issues related to the parameters and type of statistical test

In many cases, the sample size remains limited. In these situations, with, e.g. less than 20 subjects per group one can do the following (Stevens, 1986): 1) Adopt a more lenient α level, perhaps $\alpha = 0.10$ or $\alpha = 0.15$, or 2) Use one-tailed tests where the literature supports a directional hypothesis. The subsequent test will be more powerful if the experimental results are in the predicted direction (Cohen, 1969).

3.3. Reducing the within-group variability

Large within-group variability is often the reason for low power. Therefore, considering ways of reducing within-group variability is beneficial to obtain a more sensitive design (Fisher, 1934; Stevens, 1986). For example, in the sample selection more homogenous subjects can be used as they vary less on the dependent variables. One should also ensure that there is a strong linkage between the independent variables and the dependent

variable(s), and that the independent variables extend over a long enough period of time to produce a large or at least fairly large effect size.

Within-group variability can be also reduced by the experimental design (Fisher, 1934; Stevens, 1986). This can be accomplished, e.g. by using many kinds of factorial designs, analysis of covariance (covary out the confounder) or block designs (e.g. in randomized block design, the data are divided into more homogenous subgroups). Covariates, i.e. independent variables, which have low correlations with each other, are particularly useful. One can also use repeated measures designs. These designs are particularly useful, because all individual differences due to the average response of subjects are removed from the error term; individual differences contribute most of the within-group variability. Clinical cross-over trials are an example of a controlled repeated measures design (Pocock, 1993). In the most common type of cross-over trial, each patient receives two treatments one after the other, and the order of treatments is determined randomly. However, cross-over studies are not suitable for examining all kinds of diseases (see Pocock, 1993, p. 110). In addition, to cross-over patients from one treatment to another is generally not straightforward. There can be, for example, carry-over effects where the effect of the first treatment is propagated to the other treatment.

3.4. Reducing the measurement error

Random measurement error, be it due to unreliability, observational carelessness, dirty test tubes, or any other source, reduces the precision of sample results, and thus reduces power. This is because the variability of the observations is increased beyond their necessary “true” variability (Cohen, 1969). The reduction of the random measurement error will, therefore, increase the statistical power. In general, anything which reduces the variability of the observations by the exclusion of sources of variability which are irrelevant to the assessment of the phenomenon under study, will serve to increase power (Cohen, 1969). Reducing the measurement error to increase statistical power is a key issue in this thesis. We will focus on this further when we deal with the specific applications in PET.

3.5. More efficient statistical tests

The probability of the detection of the effects can be enhanced by using more sensitive techniques of analysis, i.e. more efficient statistical tests. Different FWE thresholding methods may also have an impact on power in neuroimaging (Nichols and Hayasaka, 2003). This relates to the correction for multiple tests. For example, the simplest and most widely used Bonferroni correction, based just on the number of simultaneous tests, is known to yield more conservative results (greater p values) than many of its competitors (Nichols and Hayasaka, 2003).

It should also be noted that if the distributional assumptions are not met, parametric statistical tests may not perform efficiently (or not even validly). These situations include, e.g. outlying observations or heavy tailed noise distributions. In such situations, nonparametric robust methods will yield higher efficiency (Hampel et al., 1986). This may also imply that the sample mean is not an especially sensitive measure for central tendency, and thus the use of a rank-order nonparametric test improves the statistical power (for application in neuroimaging, see Rorden et al., 2007). Similarly, robust parametric statistical methods will perform better in the presence of many outliers than ordinary parametric statistics. The presence of outliers influences the probability that the test will reject the null hypothesis (Hampel et al., 1986) and thus power. A transformation (e.g. logarithmic) can also be used to get the data to conform to normal distribution and subsequently use parametric statistics.

Chapter 4 Validation measures and their relation to power

In all the publications of this thesis, we have used measures for validation of the methods, mostly in a test-retest context. We explore the relation of these applied measures to statistical power. We also introduce also measures for validation, but these are not addressed regarding power. Before dealing with the measures, we will briefly discuss validation and introduce the additive error model of measurements.

4.1. Reliability theory: The additive error model

We will briefly describe the additive error model that belongs to the classical reliability theory of measurements (see Shrout, 1998). It should be noted that the theory described here does not deal with the reliability theory of components or systems. We assume that there is a population of measures or measurement devices, and we sample measure *j* to evaluate a fixed person *i*. We denote that measurement X_{ij} and treat is as a random variable. The measure *j* might be, e.g., an expert rater. We express X_{ij} as the sum of a person parameter, ξ_i, and a residual term carrying the unique effect of the measure *j*

$$X_{ij} = \xi_i + \varepsilon_{ij}. \tag{4}$$

The fixed person parameter, ξ_i, is defined as the expected value of *X* over the population of measures. There are various terms for the person parameter, such as the consensus score, true score and universe score. As ξ_i is $E(X_{ij})$ for a fixed person, we get $E(\varepsilon_{ij})=0$. The variance for the measurement X_{ij} is equal to $\mathrm{Var}(\varepsilon_{ij})=\sigma^2(\varepsilon)$, which is known as the squared standard error of measurement for person *i* . The smaller the $\sigma^2(\varepsilon)$ is, the more precise the measurement. The main objective in this thesis is to make $\sigma^2(\varepsilon)$ smaller and subsequently increase power.

4.2. Validation: general issues in reproducibility, reliability and validity

There are three major issues in validation of measurement methods: reproducibility, reliability and validity. All these aspects should be considered, but there is ambiguity in the literature, especially about which measures describe reproducibility and reliability. There is also obscurity about the terms reproducibility and repeatability. Repeatability is generally defined as the variation in measurements taken by a single person or instrument on the same item and under the same conditions. On the other hand, reproducibility refers to the ability of a test or experiment to be accurately reproduced, or replicated, by someone else working independently. Despite the fact that, according to the original description, we are here interested in repeatability, we will here use the term reproducibility, owing to its widespread use in the functional imaging community. Reproducibility is often called precision (International Organization of Standardization, 1993; Fig. 5). The difference between reproducibility and reliability in test-retest studies is that reproducibility deals only with (average) within subject variation, whereas reliability also takes between subjects variation into account (see Laruelle, 1999). In this thesis, we deal only with test-retest reproducibility, whereas concerning reliability, we briefly introduce also other forms of reliability. In general, reliability is defined as the consistency of a test or measurement (Weir, 2005). Validity is, generally, the ability of the measurement tool to reflect what it is designed to measure (Atkinson and Nevill, 1998). A measure may be reliable but not valid, but it cannot be valid without being reliable (see, e.g. Trochim, 2006). That is, reliability is a necessary but not sufficient condition for validity. A result is called valid if it is accurate and precise, i.e., unbiased and reproducible. In the fields of science, engineering, industry and statistics, accuracy is the degree of conformity of a measured or calculated quantity to its actual (true) value, thus being similar to bias (Fig. 5). Bias for an estimator $\hat{\mu}$ is formally defined as

$$bias(\hat{\mu}) = E(\hat{\mu}) - \mu, \tag{5}$$

where μ is the true value of the parameter. However, one should note that accuracy is a qualitative term whereas bias is a quantitative term.

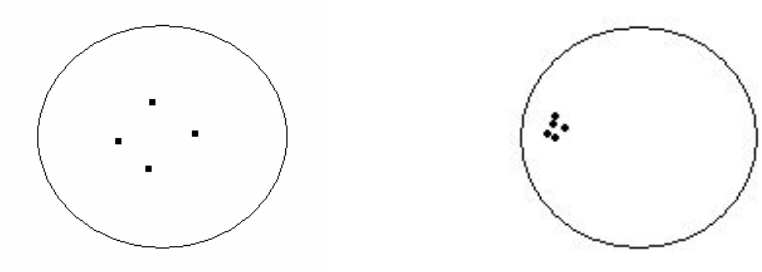

Figure 5. Target analogy: accuracy versus precision. On the left there is fairly good accuracy (lack of bias) but poor precision (reproducibility). On the right, we find high precision yet low accuracy.

4.3. Test-retest reproducibility

The term test-retest refers to two consecutive measurements of the same subject measured at different time points. There are two types of test-retest reproducibility: change in the mean and within subject variation (Hopkins, 2000). Note that Hopkins (2000) is considering reliability, whereas we regard the test-retest quantities in question as reproducibility, owing to the focus on within subject variability (see Section 4.2). Change in the mean is composed of a systematic change (bias) and a random change (variance). The random change in the mean is due to sampling error that decreases with increasing sample size. Within subject variation is important for the researchers as it affects the precision of estimates of change in the variable of an experimental study (Hopkins, 2000). The smaller the within subject variation, the easier it will be to detect or measure a change. Usually the within subject variation is expressed as the standard deviation of repeated measurements on the same subject, which enables the measurement of the size of the measurement error (Bland and Altman, 1996a). In applied sciences, measurement error is sometimes characterized as the typical error (Hopkins, 2000). Measurement error itself is generally composed of natural biological variation in the subject and variation in the measurement process, i.e. methodological error (Bland and

Altman, 1996a, Snedecor and Cochran, 1967, p.164-165). We emphasize that the reduction of methodological error is the focus of this thesis, to increase the probability of the detection of intervention effects.

4.3.1. Normalized absolute difference and its relation to power

In Publications II, III and IV, we have used a measure for test-retest reproducibility that has not been used earlier. However, closely related methods have been applied. The name of the measure is normalized absolute difference (NAD), defined as

$$NAD = \frac{|X_{i1} - X_{i2}|}{X_{i1}}, \tag{6}$$

where X_{i1} is positive. NAD, expressed as percentages and averages over the samples, has a very close relation to receptor occupancy (OCC) defined as

$$OCC = \frac{X_{i1} - X_{i2}}{X_{i1}}. \tag{7}$$

In Eq. 6, both measurements are performed at baseline, but in Eq. 7, the first measurement is performed at baseline and the second during intervention (e.g. drug). OCC is often expressed as percentages. X is typically the neurobiologically interesting parameter, binding potential (BP; see Chapter 5). NAD can be regarded as absolute variation of receptor occupancy in a test-retest setting, or a percentage signal change. Using average NAD and, e.g. 95% confidence interval for proportions, it can be inferred how accurately receptor occupancy can be computed in reality, i.e. taking test-retest variability into account. We have used the absolute sign in NAD to keep the reproducibility measure positive. The rationale for this is that regarding receptor occupancy, we can assume that the value at baseline is greater than the value during intervention (drug) as the drug occupies the receptors (see Chapter 5). NAD without the

absolute sign has been used by some authors (e.g. Hirvonen et al., 2003; Kim et al., 2006).

The relation of NAD to power is not straightforward, as the test-retest measurements are dependent on each other. However, relation to power can be examined by investigating the effect size in a paired t test using Eq. 3 (e.g. we can compare the two methods presented in Publication II in terms of the effect size, and see empirically whether improved reproducibility coincides with smaller effect size). A small effect size here is beneficial, as there is no true intervention. It should also be noted that plotting the absolute difference and mean is equivalent to plotting the standard deviation of the difference and mean (Bland and Altman, 1996b). This gives a relation to power for reproducibility, but not exactly for NAD as in NAD the first value is used in the denominator instead of the mean.

4.3.2. Variance and its relation to power

Variance measures the test-retest reproducibility also in simulations with dynamic numerical phantoms as well; see Publication I, where there was no biological variability involved. Variance in a difference variable of two independent variables has a direct connection to power in a paired t test through standard deviation (square root of variance). This is evident from the equation of effect size (Eq. 3; Section 2.4).

4.3.3. Other reproducibility measures used in functional neuroimaging

There is a large variety of other reproducibility measures used in the context of functional neuroimaging.

Coefficient of variation (CV) is defined as the sample standard deviation divided by the sample mean (Eq. 8)

$$CV = \frac{S}{\bar{X}}. \tag{8}$$

CV describes the typical error, i.e. the standard deviation of an individual's repeated measurements (Griffin, 1962, p. 116, Hopkins, 2000). We define CV as measure reproducibility, if it concerns within subject measures. If between subjects measures are considered, CV is a measure of absolute reliability (consistency of scores of individuals).

Test-retest variability (TRV; Parsey et al., 2000) is expressed as

$$TRV = \frac{|X_{i1} - X_{i2}|}{(X_{i1} + X_{i2})/2}. \quad (9)$$

The difference between NAD and TRV is small, if the test and retest values are close to each other, which is the case in PET neuroreceptor studies.

Repeatability coefficient (RC; British standard institution, 1979) is defined as

$$RC = 2S(X_{i1} - X_{i2}). \quad (10)$$

Of the differences, 95% are expected to be less than the repeatability coefficient (Bland and Altman, 1986). This is because the coefficient 2 in Eq. 10 is very near the 95% critical value, 1.96, of the normal distribution.

The average parameter value (APV) is a measure of the coefficient of variation of the difference between methods:

$$APV = \frac{2S(X_{i1} - X_{i2})}{(X_{i1} + X_{i2})/2}. \quad (11)$$

To facilitate comparisons across the regions of interest, RC is calculated as a percentage of the mean to obtain the APV (Tauscher et al., 2001; Chow et al., 2007). Those studies call APV the RM%, but the abbreviation RM is not explained there.

Root mean squared error (RMSE) is a combination of bias and variance (see Publication I). It is a square root of the mean squared error (MSE) defined for an estimator $\hat{\mu}$ to estimate the population parameter μ as

$$MSE(\hat{\mu}) = E((\hat{\mu} - \mu)^2). \tag{12}$$

Thus RMSE is

$$RMSE(\hat{\mu}) = \sqrt{MSE(\hat{\mu})} = \sqrt{Var(\hat{\mu}) + (bias(\hat{\mu}, \mu))^2} \ . \tag{13}$$

In simulations, it is often useful to calculate the sample mean of RMSE

$$\hat{RMSE}(\hat{\mu}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{\mu}_j - \mu)^2} \ , \tag{14}$$

where $\hat{\mu}_j$:s are the n realizations of the estimator $\hat{\mu}$. In RMSE, bias refers to "change in the mean" type of reproducibility, or internal validity, and variance in simulations, e.g. of the kind in Publication I, the test-retest reproducibility. Therefore, we categorize RMSE here as a measure of reproducibility. However, RMSE is a general error measure applied extensively, e.g. in simulations, where MSE and RMSE are most useful. This is because the true value of the parameter has to be known.

4.4. Reliability

There are various forms of reliability; e.g. test-retest, intra-rater, inter-rater and internal consistency. The main emphasis in this thesis is on the test-retest reliability.

In classical reliability theory, the reliability coefficient (R_x) is defined as the ratio of the true score variance to the total variance of the observed measure, X (Shrout, 1998). The observed variance of X (σ_X^2) can be formulated as the sum of the true score ($\sigma^2(\xi)$) and error ($\sigma^2(\varepsilon)$) variances

$$R_X = \frac{\sigma^2(\xi)}{\sigma_X^2} = \frac{\sigma^2(\xi)}{[\sigma^2(\xi) + \sigma^2(\varepsilon)]}. \quad (15)$$

The reliability coefficient is interpreted as the proportion of σ_X^2 that is due to replicable differences in subjects. Another equivalent formulation by, e.g. Bartko (1966) and Baumgartner (1969) is

$$R_X = \frac{\text{between subjects variability}}{\text{between subjects variability + error variance}}. \quad (16)$$

In Eq. 16, the between subjects variability, also known as between subjects mean sum of squares (BSMSS), is expressed in studies with repeated observations as:

$$BSMSS = \frac{\sum_{n=1}^{N} K(\bar{X}_{n.} - \bar{X}_{..})^2}{N-1}, \quad (17)$$

where K is the number of repeated observations per subject, $\bar{X}_{n.}$ is the mean of observations for subject n, and $\bar{X}_{..}$ is the mean of all observations. Correspondingly, the within subject variability, i.e. within subject mean sum of squares (WSMSS) is written as

$$WSMSS = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K}(X_{nk} - \bar{X}_{n.})^2}{N(K-1)}, \quad (18)$$

where X_{nk} is the k:th observation of a subject.

4.4.1. Test-retest reliability

A common form of test-retest reliability is retest correlation. The correlation coefficients indicate the reproducibility of the rank order of subjects on retest. This is called the relative reliability (Weir, 2005). On the contrary, absolute reliability concerns the consistency of scores of individuals (Weir, 2005).

There is an important difference between measurement error and retest correlation. Measurement error can be estimated from a sample of subjects that is not particularly representative of the population one wants to study (Hopkins, 2000). On the other hand, the retest correlation depends on the way the sample is chosen (Bland and Altman, 1996b).

4.4.1.1. Intra-class correlation coefficient

Intraclass correlation coefficient (ICC) was introduced in its current form by Shrout and Fleiss (1979). There are three main forms of ICC, depending on the type of reliability study. In general, ICC concerns inter-rater reliability: each of a random sample of N targets is rated independently by K judges. In the first form, each target (test-retest: subjects) is rated by a different set of K judges, randomly selected from a larger population of judges, i.e. in the test-retest studies the repeated observations in time. In the second form, a random sample of K judges is selected from a larger population, and each judge rates each N target. On the contrary, the third case concerns the rating of each target by each of the same K judges, who are the only judges of interest. In this thesis, we have used exclusively the first form, denoted in Shrout and Fleiss (1979) by ICC(1,1). It is based on the one-way random effects analysis of variance model, and is the only suitable form of ICC for test-retest studies (K = 2):

$$ICC(1,1) = \frac{BSMSS - WSMSS}{BSMSS + (K-1)WSMSS}. \qquad (19)$$

The general form of the ICC is a ratio of variance due to differences between subjects (the signal) to the total variability in the data (the noise) (Weir, 2005). ICC measures the relative reliability (see Section 4.4.1).

The increase in ICC will result in increased power to detect the effects (Perkins et al., 2000; Charter, 1997). To assess how much an increase in ICC will increase power, see e.g. Figure 2 of Perkins et al. (2000). As a rule, the method with highest the ICC should be chosen (Laruelle, 1999). A drawback of ICC is that its value depends on the population from which the study subjects have been drawn, and this may lead to difficulties in comparing results from different studies (see Section 4.4.1). The advantage of ICC is that it is also possible to use it in more than two repeated measures per subject (Atkinson and Nevill, 1998). In general, ICC is a coefficient that ranges from -1 (no reliability; BSMSS = 0) to 1 (maximum reliability; WSMSS = 0) (Parsey et al., 2000). ICC also measures the relative homogeneity within groups relative to the total variation (Kelder et al., 1993; Publication II). An ICC value of at least 0.7 is regarded as acceptable reliability (Hripcsak and Heitjan, 2002). This is, however, arbitrary.

4.4.1.2. Pearson's product moment correlation coefficient

Pearson's product moment correlation coefficient (r) was introduced by Pearson (1896). This most common measure of correlation is not typically recommended to test the test-retest reliability; however, the recommendation is not universal (Rousson et al., 2002). The main reason for the difference between r and ICC is the treatment of systematic error, which is often due to a learning effect on the test-retest data. It seems that ICC(1,1) applied in the publications of this thesis does not include a term for systematic error. In these circumstances, with two measurements made on the same subjects, the usual r is equivalent to ICC, with a slight restriction (Rousson et al., 2002). The restriction is that with r, the orders of any pairs affect the results, whereas ICC estimates the average correlation between all possible pairs of observations (Bland and Altman, 1996b). Pearson's r cannot be used with more than two repeated measurements per subject (Atkinson and Nevill, 1998).

4.4.1.3. Other test-retest reliability measures

There is a retest correlation measure that slightly differs from ICC: **the reliability coefficient (ReC**; Scheffe, 1959; Chan et al., 1998). The nominator in ReC does not have a within subject term:

$$\mathrm{Re}\,C = \frac{BSMSS}{BSMSS + WSMSS}. \tag{20}$$

This measure can be regarded as a sample version of Eq. 15. The change in ICC will reflect a change in ReC in the same direction, but with a different magnitude.

There is a widely used measure called **limits of agreement (LOA)** (Altman and Bland, 1983; Bland and Altman, 1986), which is defined for differences as

$$LOA = \bar{\bar{X}} \pm 2S. \tag{21}$$

This measure has been developed to examine the agreement between two different techniques of quantifying a certain variable. It has also been cited as a measure of absolute reliability (Atkinson and Nevill, 1998). The use of LOA as an index of reliability has been recognized (e.g. Atkinson and Nevill, 1998) but also criticized (e.g. Hopkins, 2000), while Rousson et al. (2002) states that assessing reliability is different from comparing measurement methods. However, the assessment of these two notions has common aspects (Rousson et al., 2002).

A further measure for measuring test-retest reliability is the **standard error of measurements (SEM)** (Weir, 2005). It is related to ICC as follows:

$$SEM = S\sqrt{1 - ICC}. \tag{22}$$

SEM is a measure of absolute reliability. Due to this property, SEM is more appropriate for comparing reliability between different measurement tools in different studies than ICC that indicates the relative reliability.

4.4.2. Intra- and inter-rater reliability

Intra-rater reliability is used to assess to what degree one rater gives consistent estimates of the same phenomenon, whereas inter-rater reliability concerns the same concept with multiple raters or observers.

4.4.2.1. ICC and Pearson's r

In continuous measurements, intra- and inter-rater reliability are quantified typically with ICC, as introduced earlier (see Section 4.4.1.1. for different types of rater reliability). Pearson's r is also sometimes used to assess rater reliability. However, caution must be used in the interpretation of r, as it is unaffected by the presence of any systematic biases (Hunt, 1986). A t test of the significance of r reveals whether inter-rater means differ, i.e. the presence of bias.

4.4.2.2. Kappa measures for categorical measurements

In categorical measurements, Cohen's Kappa (κ; Cohen, 1960) is suitable for assessing the inter-rater reliability of two raters. The equation for κ is

$$\kappa = \frac{\mathrm{P}(a) - \mathrm{P}(e)}{1 - \mathrm{P}(e)}, \tag{23}$$

where P(a) is the relative observed agreement among raters, and P(e) is the probability that the agreement is due to chance. Fleiss's Kappa (Fleiss, 1971) is an extension of Cohen's kappa for more than two raters. It is equivalent to ICC (e.g. Rae, 1988).

4.4.3. Internal consistency

Internal consistency can be formulated as the extent to which tests or procedures assess the same skill, characteristic or quality. There are many forms of internal consistency: average inter-item correlation, average item total correlation and split-half reliability (Trochim, 2006). The average inter-item correlation uses all the items on our instrument that are designed to measure the same construct. The average inter-item correlation is simply the average or mean of all these correlations. The average item total correlation also uses the inter-item correlations. In addition, a total score for all items is computed and used as a variable in the analysis. In split-half reliability, all items intended to measure the same construct, are randomly divided into two sets. All items are administered to a sample of people, and the total score is calculated for each randomly divided half.

4.4.3.1. Cronbach's alpha

Cronbach's Alpha (α; Cronbach, 1951) is mathematically equivalent to the average of all possible split-half estimates, and also equivalent to one form of ICC, namely ICC(3,k), defined in Shrout and Fleiss (1979). It is a classical measure for internal consistency type reliability, and is based on earlier works by Spearman (1910) and Brown (1910) as well as Kuder and Richardson (1937). It has become a universal procedure for estimating reliability and is defined as

$$\alpha = \frac{N}{N-1}\left(\frac{\sigma_X^2 - \sum_{i=1}^{N}\sigma_{Y_i}^2}{\sigma_X^2}\right), \tag{24}$$

where N is the number of components (items or testlets), σ_X^2 is the variance of the observed total test scores, and $\sigma_{Y_i}^2$ is the variance of component i for person y. There are various modifications of Cronbach's Alpha. The estimation in internal consistency is based on the correlation among the variables comprising the set. Cronbach's Alpha was applied in Publication V in the validation of subjective ratings.

4.4.3.2. Recent measures for internal consistency

Recently, a more general, multivariate measure has been developed for internal consistency, Tarkkonen's ρ (Tarkkonen and Vehkalahti, 2005). Chronbach's α is a special case of Tarkkonen's ρ in certain one-dimensional measurement models. Another alternative measure to Cronbach's α is provided by Werts et al. (1978).

4.5. Validity

4.5.1. General issues on validity

There are multiple types of validity. Some of them are related to statistics, others are not. As we have noted above, perhaps the most important statistical terms related to measurement validity are precision (reproducibility) and accuracy (lack of bias). Sometimes also the validity of the statistical conclusion is investigated. This refers to the degree to which one's analysis allows one to make the correct decision regarding the truth or approximate truth of the null hypothesis. There are also multiple types of issues where validity is applied, namely the validity of measurement, design and sampling. Validity of measurement is called the construct validity, validity of design the internal validity, and validity of sampling the external validity. Most of the material in this subsection has been condensed from Trochim (2006).

We define here the terms concerning validity of measurement, i.e. construct validity. This type of validity refers to the inferences that can be made from the operationalizations of the study to the theoretical constructs. Construct validity involves generalizing from the programme or measures to the concept of the programme or measures. The types of construct validity are translation validity and criteria-related validity. In translation validity, an attempt is made to assess the degree to which the construct can be accurately translated into the operationalization. In criteria-related validity, the performance of operationalization is checked against some criteria. Translation validity and criteria-related validity are further divided into subcategories. Translation validity comprises face validity and content validity. Face validity involves investigating whether the

operationalization seems to be a good translation of the construct. Content validity, on the other hand, is related to checking the operationalization against the relevant content domain for the construct. Criteria-related validity is composed of predictive validity, concurrent validity, convergent validity and discriminant validity. In predictive validity**,** we assess the operationalization's ability to predict something which it should theoretically be able to predict. Concurrent validity refers to the operationalization's ability to distinguish between groups that it should theoretically be able to distinguish between. Convergent validity describes the degree to which the operationalization is similar to other operationalizations that it theoretically should be similar to. Finally, the discriminant validity refers to the degree to which the operationalization is not similar to other operationalizations that it theoretically should not be similar to.

The validity of design is called the internal validity. It can be characterized as the degree to which the results of an experimental method lead to clear-cut conclusions. Internal validity is the approximate truth about inferences regarding cause-effect or causal relationships, and thus is relevant only in studies that try to establish a causal relationship. Internal validity refers to biases in statistical design. These biases can be reduced by random assignment of subjects, or by using counterbalancing for interfering variables. Internal validity can be also threatened by experimenter bias. There are also other threats to internal validity.

The validity of sampling can be characterized as external validity. This involves the extent to which the result can be generalized beyond the sample. Thus, external validity refers to the approximate truth of conclusions that involve generalizations. Sometimes, external validity is further divided into population validity and ecological validity. Population validity refers to the extent to which the findings can be generalized to other populations of people. On the contrary, ecological validity (mentioned in Publication V) is present when the experimental procedures resemble real-world conditions. Some authors think that ecological validity does not belong to external validity, because they are independent: a study may be externally valid but not ecologically valid and vice versa (Brewer, 2000; Shadish et al., 2002). It has also been noted that ecological validity is not

necessary for the overall validity of an experiment, unlike internal and external validity (Shadish et al., 2002).

4.5.2. Bias

Bias (Eq. 5, Section 4.2) measures the quantitative accuracy of measurement in validity, or assesses the systematic change in reproducibility. It also refers to internal validity (experimenter bias). In this thesis, we have used average bias over the noise realizations in a simulation study (Publication I):

$$average\ bias = g(\hat{\mu}_i - \mu), \tag{25}$$

where g(.) was either the sample mean or the sample median. Bias is useful only in simulations, because the true value of the parameter has to be known. Bias has no direct connection to power as such. It affects power only if bias is different in different classes of independent variable(s). This may be true in some circumstances in neuroimaging, so care must be taken in interpretation of results.

4.5.3. Positive predictive value

Positive predictive value (PPV) is a validity measure in a binary classification test (Pepe, 2003; Publication II). It refers to criteria-related validity of measurements. PPV is defined as

$$PPV = \frac{|TP|}{|TP| + |FP|}, \tag{26}$$

where |TP| (|FP|) is the number of true positives (false positives). In this thesis, we have used PPV in the evaluation of validity of the automatic and manual region of interest segmentation methods in PET. In this context, the use of PPV (and the Jaccard coefficient, Section 4.5.4) requires that the other set is "ground truth", as is the case with

the phantom simulations. PPV was the most appropriate measure of validity for our purposes (see Publication II for details). The relation of PPV to power can be seen in Table 1 (p. 17, Section 2.3.4).

4.5.4. Jaccard coefficient

Another measure for testing validity is the **Jaccard coefficient** (see, e.g. Jackson et al., 1989). This measure (Eq. 27) is also known as the Tanimoto coefficient and is expressed as the ratio between the size of the intersection and the size of the union of sets

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{27}$$

A and B are binary volumes. The Jaccard coefficient is generally a similarity coefficient that can, in some circumstances, also be applied to assess validity.

Chapter 5 Positron emission tomography – issues on statistical power

5.1. PET – Introduction

PET is a functional medical imaging modality. This means that with PET imaging, images can be obtained that depict some biologically or medically relevant function (Fig. 6), not anatomy, unlike, e.g. in magnetic resonance imaging. With PET, it is possible to quantitatively measure biochemical and physiological processes in vivo using radiopharmaceuticals labelled with positron-emitting radionuclides, i.e. tracers (Paans et al., 2002). After injecting these radionuclides into the body, the regional concentration of the labelled compound can be imaged as a function of time (Cherry, 2001). To obtain a dynamic image with multiple time frames and thereafter calculate corresponding time-activity curves in particular regions (Fig. 7), image reconstruction has to be carried out. This is because the measurements obtained from PET scanners are in a sinogram matrix containing the measured projections. Imaging data are then combined with a measure of the time course of the plasma probe concentration, which reflects its delivery to tissue. Thereafter, the data are processed with a compartmental model that includes equations describing the transport and reaction processes the probe undergoes (Phelps, 2000).

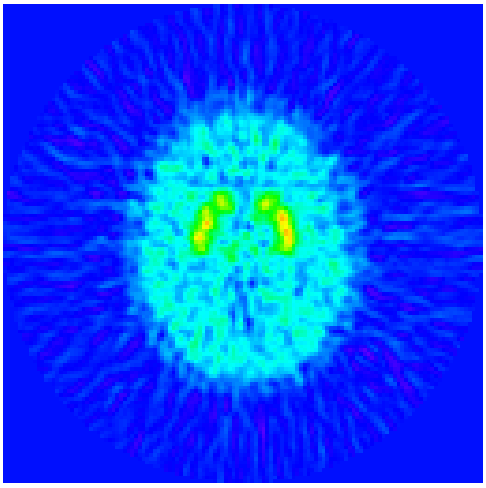

Figure 6. A sample example of a dynamic PET image. The transaxial slice is shown in the middle of data acquisition (time frame 7 of 13). The PET tracer is [^{11}C]raclopride. Picture courtesy of Turku PET Centre.

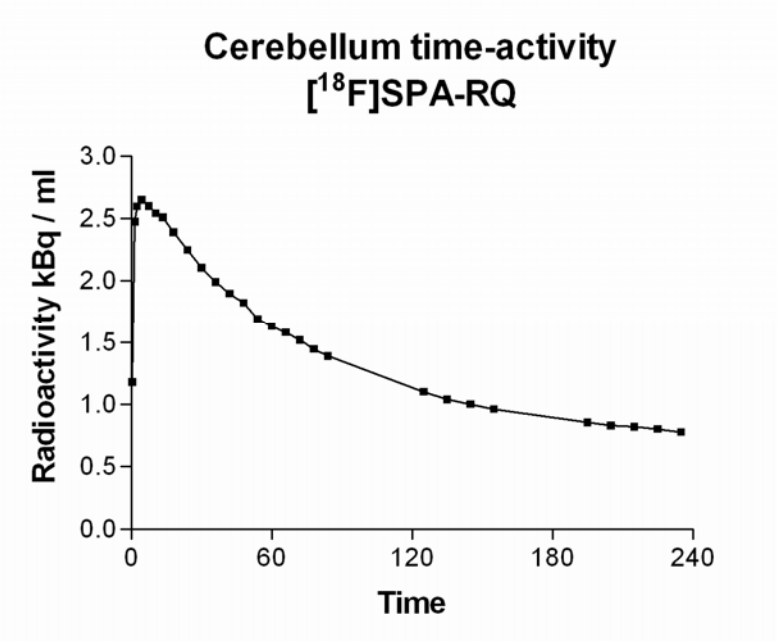

Figure 7. PET time-radioactivity curve in brain region cerebellum. The unit of activity is kiloBecquerel per milliliter (kBq /ml) and the PET tracer is [^{18}F]SPA-RQ. The data for the curve were obtained from Turku PET Centre.

PET is being used in various applications such as studies of blood flow and glucose metabolism in the heart and brain, receptor systems in the brain, oncology in whole-body scans, and recently also reporter gene expression. PET is also recognized as an important clinical tool, e.g. in cancer, cardiovascular disease, neurodegenerative disease, epilepsy, and brain injury (Cherry, 2001). Earlier, PET was also used in studying cognition, but now these studies are mostly performed with functional magnetic resonance imaging (D'Esposito, 2000). The most important application of PET concerning this thesis is, however, early drug development. PET offers possibilities to accelerate the drug development process from preclinical discovery to phase III clinical trials (Eckelman, 2002). PET is being used in preclinical and phase I clinical trials mostly in the central neural system (psychiatry), in anti-cancer (oncology) and small-animal PET-related drug development (Lee and Farde, 2006; Saleem et al., 2006; Cherry, 2001). In this thesis, we deal specifically with methodological properties of receptor occupancy (Fig. 8), which guides the dose-finding procedures for a drug (see, e.g. Passchier et al., 2002 for review). In receptor occupancy, there is a competition between the drug and the PET tracer. Occupancy gives the proportion of receptors that the drug has occupied, thus defeating the tracer. In conclusion, the advantages of PET are the excellent sensitivity, flexible chemistry and decay scheme that lead to favourable imaging properties. The drawbacks include cost and accessibility (Cherry, 2001).

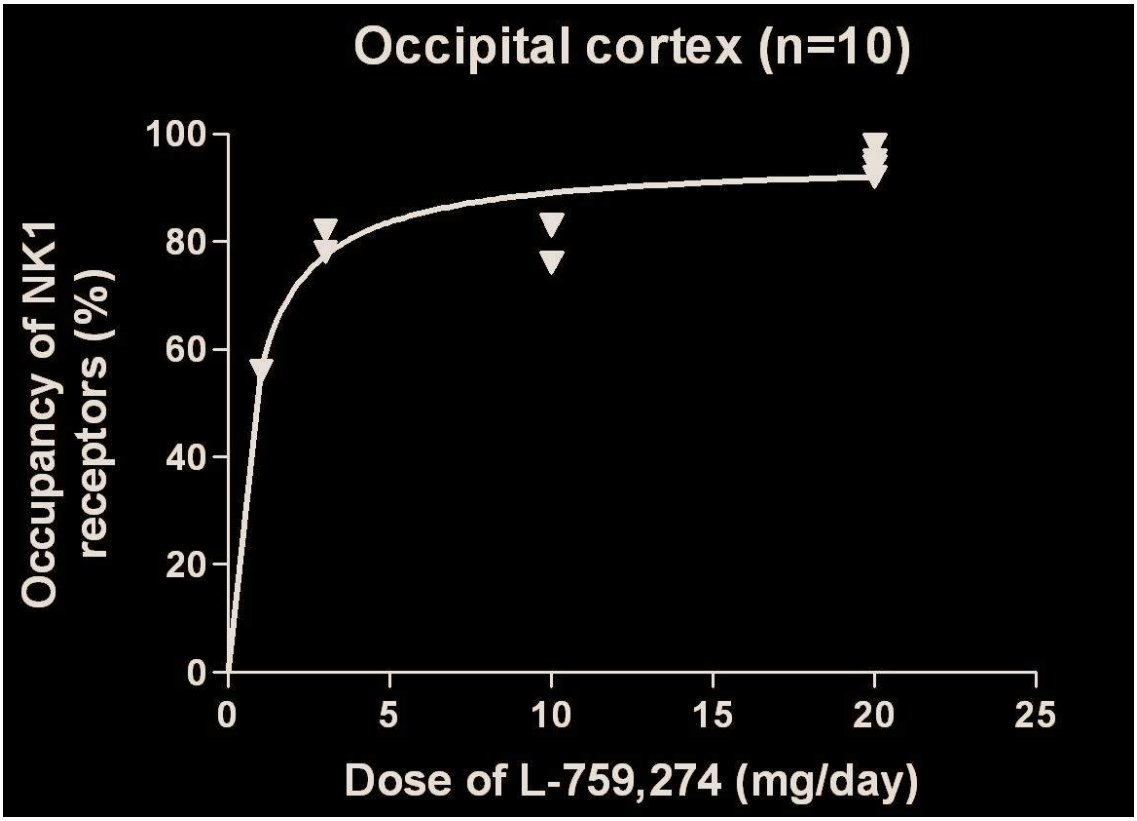

Figure 8. Example of receptor occupancy % (vertical axis) as a function of the dose of a drug L-759-274 (horizontal axis) in the occipital cortex of the brain. The PET tracer is [^{18}F]SPA-RQ, which binds to NK1 receptors. Picture courtesy of Turku PET Centre.

With real PET data, it is impossible to separate the biological and methodological sources of variation and errors in the images. Therefore, real PET data are not suitable for testing different sources of variation and errors. To simulate these, both numerical and physical phantom images have been developed, i.e. artificial PET images. The phantom simulations are also very important in developing and testing models. In this thesis, we dealt with dynamic numerical phantoms. In Publication I, we first investigated the dynamic behaviour of the PET tracer by utilizing values from real PET studies. Accordingly, there, we constructed the virtual radioactivity distribution on the chosen anatomy. We were then able to imitate for the first time the whole procedure and

methodological errors of receptor occupancy studies by phantom simulations. To a limited extent, also Monte Carlo-simulated dynamic numerical phantom images (Reilhac et. al, 2004) were used in Publication I, as well as in Publications II, III and IV. The Monte Carlo-based simulator introduced in Reilhac et al. (2004) provides a realistic model for data acquisition that takes into account the system dead time, random and scatter coincidences, detector normalization, and attenuation.

5.2. Challenges in PET concerning the statistical power

There are many challenges concerning the statistical power in PET. Abundant noise appears in the measured PET projections that impair the quality of the reconstructed images (Ollinger and Fessler, 1997). The noise originates from the statistical nature of PET data acquisition: two photons are simultaneously detected yielding a "coincidence´" event in a line of response between two detectors in the PET scanner (Turkington, 2001). This means that the PET measurements are based on a counting process yielding independent Poisson distributions for the measurements (Ollinger and Fessler, 1997). Moreover, the quality of PET images is degraded by various physical factors, some of which can be corrected. The degrading factors include scatter, attenuation, random events (accidental coincidences), detector dead time, and detector normalization (see Turkington, 2001 for descriptions and details). In the statistical sense, the most important factor is random events, the current correction of which renders the data non-Poisson, with variance greater than the mean (Fessler, 1994, Yavuz and Fessler, 1998). With PET, there is also an ethical limiting factor: the radiation dose to the study subjects. This radiation restricts increasing the study sample size (with a few exceptions). Therefore, many PET studies suffer from low statistical power. Further factors that impair statistical power in PET studies can be easily found. There are, namely, many preprocessing steps before the images to be analyzed can be obtained. In general, the variation in PET studies consists of inherent biological variation in the populations under study, and the preprocessing steps: errors due to instrumentation and image reconstruction, physiological quantification inaccuracies, and errors in extraction of regional information

from the imaging data (Carson, 1991). Except for biological variation, these errors can be classified as errors arising from the applied methods. As the methodological errors can be considerable, they may have a substantial effect on statistical power.

5.3. Statistical analysis of PET image data

5.3.1. Preprocessing of PET image data for statistical analysis

Roughly speaking, there are two distinct ways to statistically analyze brain PET image data: region of interest (ROI) -based analysis and voxel-based analysis (a voxel is a three-dimensional volume element). Actually, the ROI-based analysis concerns three-dimensional volumes of interest (VOIs) instead of regions, but we will use the conventional name ROI. Before the data can be statistically analyzed, various preprocessing steps have to be performed.

The PET measurements are acquired in sinogram matrices that cannot yet be analyzed as such, although some noise reduction in them has been performed through radial (see, e.g. Kao et al., 1998; La Riviére and Pan, 2000) and sinusoidal trajectory based filtering (Happonen and Alenius, 2005; Peltonen and Ruotsalainen, 2006) . An image is created from sinogram matrices using image reconstruction. There are, roughly, two types of reconstruction methods, analytical (Jain, 1989) and statistical iterative reconstructions (e.g. Yavuz and Fessler 1998; Alenius and Ruotsalainen, 2002). Thereafter, the obtained dynamic emission image is computed through compartmental modelling to a parametric brain image (voxel-based analysis) or parametric values in specific brain areas of interest (ROI-based analysis), using the time-activity values. The parametric images or values can be obtained using various methods (see, e.g. Logan et al., 1996; Gunn et al., 1997; Cselényi et. al., 2006), some of which cannot be used with every PET tracer.

Usually, image registration is also done prior to analysis (e.g. realignment: same subject at different times, co-registration: same subject, different imaging modalities, spatial normalization: different subjects, same imaging modality). Co-registration between PET and magnetic resonance (MR) image is typically used in ROI-based analysis, and spatial

normalization in voxel-based analysis to fit the images into some standard stereotaxic space (Ashburner and Friston, 2005). Then, there is the extraction of ROIs (possible in both types of analyses), and spatial smoothing, typically performed only in voxel-based analysis. It should also be noted that PET quantification is hampered by partial volume effect and spillover (e.g. Rousset et al., 1998; Hoffmann 1979), which are due to limited spatial resolution (partial volume) and its reciprocal effects (spillover). Moreover, their effect on quantification is different; partial volume decreases tracer uptake, whereas spillover increases it. Due to these degrading factors, outliers appear in the PET image data. This implies that without the partial volume correction - which can be performed in some cases - the sample mean may not be adequate to estimate the central tendency. This was obvious in Publication II, where the Monte Carlo phantom results with anatomical ground truth structures provided poor mean quantification, and also in Publication I, although partial volume was not explicitly modelled there. All these causes produce errors that increase the measurement error, and thus may have even a significant effect on power. In this sense, the fewer modifications to the data to be analyzed the better. However, some modifications are required in PET, depending on the goal of the analysis.

5.3.2. ROI-based statistical analysis

The conventional ROI-based statistical analysis is based on average (e.g. mean, median) intensity values in specific ROIs (e.g. Huesman, 1984). In PET neuroreceptor studies, this is typically performed from BP values (see Hirvonen et al., 2003 for typical practical implementation). From BP values of the baseline and drug intervention scan, we can then calculate the receptor occupancy in ROIs, which depict the percentage of receptors that the drug has occupied. The use of average intensity values in a ROI means that there is only one value in a ROI to be analyzed from an image. Therefore, the data need not necessarily be presented in the image domain, thus enabling the use of standard statistical packages in the analysis. Because these can be used, the generalizability of the results of ROI-based analysis is at the same level as in typical statistical analyses. In most cases, this means that the results can be generalized from the sample to the population level, provided that the sample selection is appropriate. Due to the averaging of voxels inside a

ROI, statistical uncertainty is also reduced (Huesman, 1984). The drawback of ROI-based analysis is that it is not as spatially specific as voxel-based analysis, i.e. the differences within ROIs cannot be studied. In addition, it may not be very suitable for finding new brain areas, where significant effects might be found. In many cases, the ROIs to be analyzed are chosen on the basis of the earlier literature.

In this thesis, ROI-based statistical analysis was not used as such. However, all publications, except Publication V, use a hybrid approach in the evaluations: parametric images were calculated as in voxel-based analysis, but these were averaged in individual ROIs. This is known to produce comparable results with the conventional ROI approach with [^{11}C]raclopride tracer, and according to evaluations also with [^{18}F]SPA-RQ (evaluations were not included in the final version of Publication I). It should be noted that in these publications, conventional statistical analysis was not performed, only methodological evaluations (bias, variance, reproducibility, reliability, etc.).

5.3.3. Voxel-based statistical analysis

Compared to the ROI-based analysis, the voxel-based technique is a more recent way to analyze PET imaging data. The idea of voxel-based analysis is to examine each voxel in the entire brain simultaneously (i.e. each voxel has a null hypothesis of no effect on its own). This massively univariate approach is based on GLM that allows the residuals to have a variance that varies from voxel to voxel; otherwise residuals have the usual assumptions of normal and independent distribution with zero mean (Friston et al., 1995). There are also other strong assumptions (see Section 5.4.2). The results are called statistical parametric maps, defined as spatially extended stochastic processes that are used to test hypotheses about regionally specific effects in neuroimaging data. The whole approach is called statistical parametric mapping (SPM). There is a statistical package, SPM (e.g. Friston et al., 1995), for voxel-based analysis in functional neuroimaging, as well as various competing packages, such as FSL (Smith et al., 2004), AFNI (Saad et al., 2006) and others. Within the voxel-based analysis framework in SPM, the inference can also be based on clusters of voxels instead of single voxels, i.e. cluster-based inference can be performed (Friston et al., 1994). In this inference, the spatial extent of activated

regions is taken into account when determining the height threshold corresponding to the significance criterion. Cluster-based inference is suitable for detecting effects in different kinds of signals from voxel-based inference: cluster-based inference favours large but subtle effects, whereas voxel-based inference more easily detects peaked effects with high statistical scores (see Friston et al., 1996 for theory and details). In voxel-based analyses, the search volume can also be kept smaller than the entire brain, using the small volume correction (SVC; Worsley et al., 1996). The generalizability of SPM analyses is limited, i.e. they are fixed effects analyses rendering the studies "case studies" (Holmes and Friston, 1998). Therefore, the preceding authors developed a two-stage analysis (in functional neuroimaging, called random effects analysis) to extend the inference to the population level. In this thesis, Publication V uses a voxel-based statistical analysis, the results of which can be generalized at the population level.

In the voxel-based analysis, spatial normalization has to be applied to fit the images from different subjects into a standard space. This leads to errors in localization of effects, especially in small brain regions. The images are also typically smoothed to increase signal-to-noise ratio (which is inherently worse in voxel-based than in ROI-based analysis), and due to distributional assumptions of voxel-based analysis. This further hampers the localization of effects due to the spreading of the effects. Despite the aforementioned reasons, voxel-based analysis is, generally, more spatially specific than ROI-based analysis. In some cases, the voxel-based analysis also has limited generalizability of results. Voxel-based analysis can be considered as a more exploratory tool than ROI-based analysis. It is suitable for finding new brain areas, where effects might be found. This could be accomplished by performing a voxel-based analysis, at a low threshold, without correction for multiple tests, as a pre-phase for ROI selection and analysis (Holmes, 1994). An alternative to ROI analysis is to use voxel-based small volume correction (SVC) analysis (Worsley et al., 1996), but then the selection of ROIs should be made similarly as in ROI analysis, not directly based on the results of the analysis. It is not reasonable to "confirm" the results of ROI-based analysis by voxel-based analysis that uses SVC. The hypothesis testing is different, and they are thus alternatives or complementary to, but not confirmatory of each other.

5.4. Power of statistical analysis in PET and ways to enhance it

5.4.1. ROI-based analysis: power

Concerning the statistical power, ROI-based analysis is known to be more sensitive in detecting the effects than voxel-based analysis (Holmes, 1994). This is partially due to the fact that the correction for multiple tests corresponds to the number of chosen ROIs of simultaneous interest, and is less severe than in voxel-based analysis. Moreover, the averaging inside a ROI effectively reduces the statistical uncertainty, especially if the ROI is large. To increase the power in ROI-based analysis, standard methods can be used as such (Chapter 3). However, due to the ethical and financial restrictions in PET, the sample size cannot be substantially increased unless meta-analyses or multicentre clinical trials are carried out. There is literature on how to determine the adequate sample size and power in ROI analysis with PET (e.g. Wahl and Nahmias, 1998). It should also be noted that the partial volume effect influences the statistical analysis and power if the effect of partial volume is different in different treatments, subjects or groups. At least the spill-over effect is dependent on the magnitude of intensity values: high intensity values spread more than low ones. This is true for both types of PET statistical analyses.

5.4.2. Voxel-based analysis: power

There are multiple problems with statistical power in voxel-based statistical analysis. Firstly, because p values are arranged as images, they have a spatial interrelationship and spatial patterns of varying size (Turkheimer et al., 2004). This hampers estimation of the adequate sample size. Many previous articles on statistical power have thus used repeatability of effects as a criterion for determining the sample size instead of actual power analysis. There are various articles that deal with power in voxel-based functional neuroimaging (Grabowski et al., 1996; Andreasen et al., 1996; Van Horn et al., 1998; Petersson et al., 1999; Desmond and Glover, 2002), but all of them concern activation studies, not PET neuroreceptor studies.

Secondly, there are a large number of multiple tests in voxel-based statistical analysis (there are about 200 000 voxels in the brain). The correction for multiple tests in SPM voxel-based analysis is based on the theory of Gaussian random fields (Adler, 1981). The correction is not, however, based directly on the number of voxels, but the expectation of the Euler characteristic and the number of resolution elements that determine the p value (Worsley et al., 1996). The approximative equation is

$$P(M \geq u) \approx \sum_{D=0}^{3} R_D(SV)\rho_D(u), \tag{28}$$

where M is the global maximum, u the height threshold value, D the dimension, SV the search volume, R_D the number of resolution elements, and ρ_D the probability density function for the Euler characteristic. This correction is severe and makes the statistical power fairly low (see Nichols and Hayasaka, 2003 for comparison of various thresholding methods). It also makes many strong assumptions, especially about the smoothness of random fields. This is one of the reasons why scans from functional imaging experiments are typically smoothed with a Gaussian kernel prior to SPM analysis.

On the basis of previously described theory (Worsley et al., 1996), it is possible to restrict the inferences to a smaller volume than the entire brain using a priori hypothesis, i.e. using small volume corrected SVC analysis. It increases power in the hypothesized areas due to less severe correction for multiple tests. However, SVC is not unproblematic regarding power (Turkheimer et al., 2004), because there is a loss of power elsewhere in the brain. Another way to increase the statistical power is to use nonparametric methods in the correction for multiple tests (Holmes et al., 1996; Nichols and Holmes, 2002). This approach, called statistical nonparametric mapping (SnPM), uses a less severe permutation based correction for multiple tests, thus gaining more power (Fig. 9), and having less stringent assumptions than SPM (see Appendix of Publication V). It also enables the pooling of variance estimates over neighbouring voxels ("variance smoothing"), under the mild assertion that the true variance image is smooth, thus giving

additional degrees of freedom. Although the correction for multiple tests is nonparametric, the approach still uses the GLM in the estimation of effects. This approach is also suitable for PET receptor studies with small sample size (although Publication V is not a receptor study but a PET activation study).

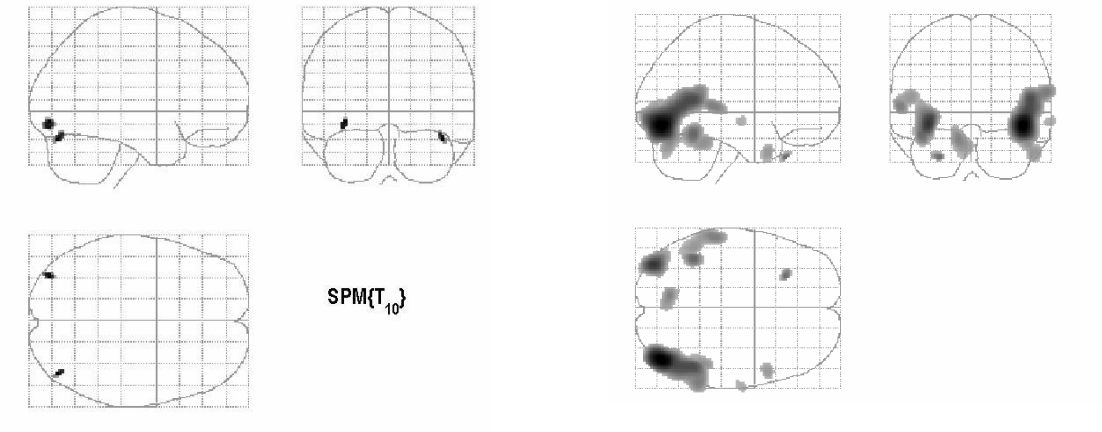

Figure 9. SPM versus SnPM results of voxel-based regression analysis in sagittal (upper row; left), coronal (upper row; right) and axial (lower row) views. The results concern responses to amusing film clips. On the left, SPM results, and on the right SnPM results. Increase in power with SnPM can be clearly seen by more statistically significant results, most of it due to variance smoothing. The significance level corrected for multiple tests was the same, 0.05, for both analyses. SnPM results are the same as in Publication V.

In addition to the nonparametric methods, the correction for multiple tests can be made more lenient by estimation of the "true" null hypotheses (Turkheimer et al., 2001). In this way, the number of tests can be reduced. A Bayesian perspective on multiple hypothesis testing can also be used (Turkheimer et al., 2004; Friston et al., 2002). In that approach, there is a separate model for the signal and the noise (Turkheimer et al., 2004). An attracting approach is to use false discovery rate (FDR) based correction for multiple tests (Genovese et al., 2002). It is adaptive and generally has better power to detect effects than, e.g. correction based on the theory of random fields. Unfortunately, the statistical inference in FDR does not obey FWE. Instead, it allows that a pre-specified proportion of statistically significant results will be false positives, but we do not know which part.

Some authors have claimed that FDR is an exploratory tool (e.g. Turkheimer et al., 2004), but this is not universally agreed.

Besides the correction for multiple tests, statistical power can be enhanced by various other analysis techniques. In the context of neuroimaging, the newest of these are truly nonparametric methods based on ranks (Rorden et al., 2007). Even in SnPM, the tests are based on mean statistics, which is not the most optimal choice in many situations, e.g. many outlying observations or skewed distributions. In such circumstances, methods based on ranks are more appropriate and have more power (Rorden et al., 2007). Another way to gain more power is to use a statistical method designed specifically for PET neuroreceptor studies (Aston et al., 2000). This method, called the residuals t test, provides increased degrees of freedom by utilizing dynamic, time-varying information from the residuals of the fit of compartmental model, and associated standard deviation images. A fundamentally different approach is provided by the wavelets-based methods in statistical analysis (Cselényi et al., 2002; Turkheimer et al., 2000). Wavelets are well known in image processing for their noise reduction properties. However, their problem lies in the statistical inference that can possibly be made implicitly (Jernigan et al., 2003).

5.4.3. Methods of this thesis concerning the power in PET statistical analysis

Our ultimate aim is to reduce confounding test-retest variation in PET by better methodology, and thus increase the statistical power to detect effects. Reducing methodological errors plays a remarkable role in the publications of this thesis (Publication I focuses on errors due to image reconstruction and physiological quantification, whereas Publications II, III and IV examine the errors derived from the extraction of regional information). Publication V, on the other hand, attempts to improve the statistical power by using a sensitive statistical technique for analysis. In other words, the first four publications try to reduce the noise, whereas Publication V tries to enhance the detection of the signal, without explicitly reducing the noise before the statistical analysis. Both approaches lead to similar results. Although Publication V concerns PET

activation studies on emotion, the same statistical method can also be used in PET neuroreceptor studies with small sample sizes.

In this section, we show in more detail the improvements made in this thesis concerning the statistical power. In a paired t test, the denominator has a standard deviation (square root of the variance) of a difference variable (Eq. 2). In Publication I it has been shown that the median root prior (MRP) (Alenius and Ruotsalainen, 1997; Alenius and Ruotsalainen, 2002) image reconstruction method has smaller standard deviation than filtered back-projection (FBP) (Jain, 1989) reconstruction in the dynamic PET images and receptor occupancy. This is the case also with BP images with both baseline and intervention scans. This implies improved statistical power, as in simulations the variance of a difference variable can be written as

$$Var(X-Y) = Var(X) + Var(Y), \tag{29}$$

where X is here the baseline scan and Y the intervention scan (Lindgren, 1968). This is because in simulations, we can assume that X and Y are independent. However, in reality, this does not hold, as X and Y are dependent due to two measurements per subject. In that case, a different measure should be used (Lindgren, 1968):

$$Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y), \tag{30}$$

where Cov(X,Y) is the covariance between X and Y. In Publication I, we also found the smallest average bias using a simplified reference tissue model with basis functions (SRTM BF; Gunn et al., 1997) for physiological quantification. Although average bias does not have an unambiguous connection to power, we recommend using MRP image reconstruction in conjunction with the SRTM BF physiological calculation method with [^{18}F]SPA-RQ tracer (Publication I). Concerning the automatic segmentation of striatum ROI used in Publications II, III and IV, we have above shown that increased reproducibility and reliability imply better power with automatic ROI extraction methods compared to manual extraction methods. Additional support for increased power in

automatic segmentation is given in Table 2, in terms of average standard deviation (S) of BP and NAD in a test-retest study (n = 8) utilized in the publications (II-IV) of this thesis.

	Pairwise DSM-OS	DSM-OS coreg	DSM-OS	manual orig	manual coreg
Average S BP	0.3465	0.3326	0.3244	0.3968	0.3558
Average S NAD	2.7646	1.8876	2.2482	3.7173	2.9315

Table 2. Average standard deviation of a reproducibility measure NAD and binding potential (BP) in brain structure striatum. Various ROI extraction methods, introduced in publications (II-IV) of this thesis, were applied. The first three columns describe the different automatic extraction methods (Pairwise DSM-OS [dual surface minimization – outer surface], DSM-OS coreg [with PET-to-PET co-registration] and DSM-OS), and the last two columns manual extraction methods (manual orig [original] and manual coreg [with PET-to-PET co-registration]).

Table 2 shows that the point estimates of average standard deviations were always smaller with the automatic methods than with the manual methods. Statistically significant differences in NAD S were found between DSM-OS coreg and manual orig ($p=0.0175$, 95 % confidence interval [0.48,3.18]) and DSM-OS coreg and manual coreg ($p=0.0453$, [0.03,2.06]). With BP S statistically significant differences were seen between DSM-OS and manual orig ($p=0.0064$,[0.03,0.11]), and between DSM-OS coreg and manual orig ($p=0.0271$,[0.01,0.11]). In calculation of the effect size, standard deviation of the difference variable should be used (Eq. 2). Concerning the voxel-based statistical analysis in Publication V, the improvement in power has already been shown (Fig. 9).

5.4.4. A proper statistical design to detect the effects of a drug with PET

A proper design for testing the effects of intervention (e.g. drug) with PET can be obtained by a 2 x 2 factorial design: a test-retest group (two baseline scans from each subject) and a drug group (a baseline scan and a drug scan from each subject). To test the effects of a drug, an interaction between replication and group is of interest. In an

appropriate test for an effect of drug, the confounding test-retest variation must be taken into account. Therefore, just testing the simple difference between baseline and drug is not acceptable. Another way to incorporate the test-retest variation in the statistical testing is to include it as a covariate while testing the effects of the intervention. A drawback in this design is that there are no replications concerning the drug scans of the same subject. This can be accomplished in the later stages of the drug development, as PET is currently used only in phase I of three-stage clinical trials.

Chapter 6 Summary of publications

Publication I

Publication I is the core of this thesis. We developed a procedure for comparing image reconstruction methods and physiological model calculation methods in PET receptor occupancy studies. However, the procedure can be used generally for examining methodological errors in PET receptor occupancy studies with any PET tracer. Average bias, variance and root mean squared error were used as error measures in a simulation study that utilized dynamic numerical phantoms. A total of eight combinations of reconstruction and physiological calculation methods were compared. The statistical iterative reconstruction method MRP produced the lowest variances, and the physiological model calculation method SRTM BF yielded the lowest average biases in receptor occupancy values. Low variance of iterative reconstruction implies greater power to detect effects in statistical analysis.

Publication II

An entirely automatic, novel method for the extraction of brain regions, the caudate and putamen, from PET BP images was introduced. We also validated the automatic method and compared it with the conventional manual extraction of the caudate and putamen, using test-retest (baseline) PET images and a Monte Carlo-simulated image. The automatic method was based on deformable surface models and normalized cuts. The global optimization-based algorithms in deformable models can be thought as applications of the algorithms presented in the PhD thesis of Tohka (2003). Those applied in the present thesis are abbreviated in Tohka's work as DSM-OS and DSM-IS. The most important part of the publication concerning this thesis is, however, the validation. The reproducibility was assessed using average NAD, the reliability using ICC, and the validity using PPV. The automatic method was found to be considerably better than manual method in terms of reproducibility, somewhat better in terms of point estimates of

reliability (indicating greater power with the automatic method), and comparable concerning validity.

Publication III

The automatic segmentation method introduced in Publication II was extended to serve the needs of receptor occupancy studies. This was achieved by assuming two scans per subject (baseline study and intervention study), and developing two modified automatic segmentation methods. The segmentation result of the baseline study was utilized in the segmentation of the intervention study, which is a challenging task due to the considerably reduced uptake of the tracer compared with the baseline study. The data consisted of real occupancy studies in addition to the data in Publication II. We validated the new methods (two variants of DSM-OS) similarly as in Publication II (with the exception that instead of validity, receptor occupancy was examined), and compared the segmentation methods with each other. Also in this case, the validation results were generally favourable for the automatic segmentation as compared to the manual one.

Publication IV

This study is complementary to Publication III. A third method for automatic segmentation of striatum was developed for receptor occupancy studies, which guide the dose-finding procedures in early drug development. This time, the deformable model-based algorithm was different (DSM-IS) from that in Publication III (DSM-OS), but the data were the same. The third automatic method was found to be comparable to the methods in Publication III and remained somewhat better than the manual segmentation, thus indicating enhanced power.

Publication V

A voxel-based subject-specific regression analysis in studies on emotion was presented. This was the first application of the method in PET studies on emotions, although the theory was developed earlier. The advantages and drawbacks of the method were discussed. However, the most important contribution of the publication to this thesis lies

elsewhere. Namely, in regression analysis, we used a voxel-based statistical method, which uses non-parametric permutation-based correction for multiple tests. It has more statistical power than parametric methods to detect effects when sample size is small (e.g. n<15). Moreover, we used a procedure enabling results that can be generalized to population level, unlike standard voxel-based analyses.

Author's contribution to the publications

The major issue of this overview and the connecting topic of the publications of this thesis – increasing the statistical power in the analysis of PET data – was the idea of E. Wallius. However, the initial ideas for the individual publications were proposed by the supervisor, U. Ruotsalainen (Publication I), in collaboration with the supervisor and the instructor J. Tohka (Publications II, III and IV), and by S. Aalto (first author of Publication V). In Publication I, the author conducted all the experiments and simulations, and wrote most of the article, while getting advice and comments from the supervisor and other authors. Publication I was a learning process; the aims of the study developed gradually in discussions between the author and the supervisor. In Publication II, the author wrote most of the Introduction, all the "Experiments with test-retest human studies" and related results, and participated in writing the Abstract, Discussion and Conclusion. In addition, the author performed all the experiments concerning the automatic segmentation in the test-retest human studies. In Publications III and IV, the author wrote the most part of the articles in guidance with the other authors, implemented the required modifications for the segmentation algorithm, and performed all the experiments. Publication V was initiated already during the author's M. Soc. Sc. thesis, and was written in tandem with S. Aalto. The main contribution and responsibility of the author was to perform the voxel-based statistical analyses and report their results, as well as write the theoretical Appendix and those parts of the Discussion that are related to statistics.

Chapter 7 Discussion

7.1. The statistical power

We have shown that the increase of power in the statistical analysis of PET imaging data can be obtained without increasing the sample size. To achieve this, we used methods broadly characterized as image analysis methods to reduce the measurement error. In Publication V, another approach, based on improved detection of effects in statistical analysis, was presented. The resulting increased statistical power benefits drug development. In the current thesis, drug development is explored through dealing with the methodological properties of receptor occupancy studies that guide the dose-finding procedures (Fig. 7). There are previous theses that have, as a minor part, dealt with statistical power in PET studies (e.g. Holmes, 1994; Reinders, 2004). Recently, test-retest studies for almost every PET tracer have also been performed, but the increased power is not typically emphasized in those studies, although it is known that the test-retest variation has an impact on power.

The finding of Publication I that the use of statistical iterative reconstruction (e.g. MRP) increases the statistical power in PET compared to analytical reconstruction methods is not new (Reinders et al., 2002; Mesina et al., 2003). However, these publications dealing with the voxel-based analysis have used a different statistical iterative image reconstruction method from this thesis, and also they are PET activation studies unlike most of the publications of this thesis. In Publication I, physiological model calculation methods were found to have an impact on average bias, which can, but does not necessarily affect power. We also found better power using completely automatic ROI extraction than the conventional manual ROI extraction (Table 2, Publications II, III, and IV). Other automatic and semi-automatic ROI extraction methods exist for PET neuroreceptor studies (e.g., Rusjan et al., 2006; Chow et al., 2007), but these rely on MR images in addition to PET images, whereas our method utilizes solely PET images. Chow et al. (2007) have previously investigated effect size in the context of semi-automated extraction of ROIs, and reproducibility is assessed in various articles.

7.2. Applications in drug development

The methods applied in this thesis may be especially useful in receptor occupancy type multicentre clinical trials and meta-analyses. In addition to the gain in power, such studies greatly benefit from automation and careful pre-planning of experiments. Normally, pre-planning also involves power analysis, which may be difficult with PET as we have discussed above. The automation of ROI extraction (Publications II, III and IV) can possibly substantially speed up early drug development together with PET. The development of new PET scanners with better resolution makes automated ROI extraction even more useful, as the burden for manual extraction increases enormously. Automation can also help to relieve intra- and inter-observer variability to increase the reproducibility (e.g. Zijdenbos et al, 1998; Ruotsalainen et al., 2001) and power. In manual segmentation, different experts have divergent opinions and individual working habits, all of which reduce reproducibility. On the other hand, pre-planning of PET receptor occupancy experiments can yield substantial cost savings, and can also improve power (Publication I). This is because the use of dynamic numerical phantom studies to select the most appropriate image reconstruction and physiological calculation methods may save many expensive pilot PET studies with humans. Possibly, time can also be saved, as the methods applied in the studies do not have to be tested with human studies but with realistic simulations. As far as we know, the best way to do this pre-planning is to use the Monte Carlo-simulations with dynamic numerical phantoms (Reilhac et al., 2004). These were to a limited extent also used in Publication I. When Publication I was in preparation, there were no test-retest studies on [^{18}F]SPA-RQ PET tracer. Recently, a test-retest study for this tracer has been published (Yasuno et al., 2007). Publication I in a way examines test-retest variability in a situation where there is only methodological variation.

7.3. Validation measures

Validation measures, especially reproducibility and reliability, as well as their relation to the statistical power, are one of the focuses of this thesis. TRV is, in general, probably a

better measure for reproducibility than the NAD we used, as the denominator includes the mean. This is because the absolute subject difference versus subject mean is equivalent to a standard deviation of the difference versus mean when there are only two measurements per subject (Bland and Altman, 1996b), thus indicating a relation to power. However, NAD suits better than TRV for examining receptor occupancy, owing to the similar form of the corresponding equations (Eq. 6 and Eq. 7). With NAD one can determine the absolute effect of test-retest variation to measured receptor occupancy (Publication II). This is valid, if we assume that the intervention (drug) BP scan always has a smaller value than the baseline BP scan. This is a reasonable assumption, as it ensures that the receptor occupancy cannot be negative. However, in proper examinations, NADs should be complemented with, e.g. 95% confidence intervals for proportions. The investigations in this thesis showed that the average NADs were in most cases under 10%. This is in line with Passchier et al. (2002), who stated that the test-retest variation in PET is 5-10%. Concerning neuroreceptor PET reliability studies, ICC has become a gold standard (Laruelle, 1999). Because we compared methods in this thesis, we could, in principle, have utilized LOA, but we decided to stick to reliability measured by ICC. With ICC, there is no learning effect in neuroreceptor PET, as BP is a quantitative measure. Thus, it is not a task that can be learned, unlike, e.g. in neuropsychology, where the systematic error (change) between test-retest experiments can cause problems. If needed for absolute reliability, ICC can be complemented with a related measure, SEM (e.g. Weir, 2005). In addition, there are no human raters in the case of automatic segmentation and usually only one trained specialist in manual extraction. But with PET test-retest experiments, the different time points at which the measurements are made can be regarded as the raters as the PET measurements are statistical in nature and vary from time to time. Of course, natural biological variation is also included. For these reasons, we conclude that with PET test-retest studies, the problems relate both to the test-retest and inter-rater reliability.

References

Adler RJ. The Geometry of Random Fields. Wiley, New York, 1981.

Aitkin M, Boys RJ, Chadwick T. Bayesian point null hypothesis testing via the posterior likelihood ratio. Statistics and Computing, 15(3): 217-230, 2005.

Alenius S and Ruotsalainen U. Bayesian image reconstruction for emission tomography based on median root prior. European Journal of Nuclear Medicine, 24: 258-265, 1997.

Alenius S and Ruotsalainen U. Generalization of median root prior reconstruction. IEEE Transactions on Medical Imaging, 21: 1413-1420, 2002.

Altman DG and Bland JM. Measurement in medicine: the analysis of method comparison studies. The Statistician, 32(3):307-317, 1983.

Andreasen NC, Ardnt S, Cizadlo T, O'Leary DS, Watkins GL, Ponto LL, Hichwa RD. Sample size and statistical power in [^{15}O]H_2O studies on human cognition. Journal of Cerebral Blood Flow and Metabolism, 16: 804-816, 1996.

Ashburner J and Friston KJ. Unified segmentation. NeuroImage, 26: 839-851, 2005.

Aston JAD, Gunn RN, Worsley KJ, Ma Y, Evans AC, Dagher A. A statistical method for the analysis of positron emission tomography neuroreceptor ligand data. Neuroimage 12:245-256, 2000.

Atkinson G and Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Medicine, 26(4):217-238, 1998.

Barron A, Rissanen J, Yu B. The minimum description length principle in coding and modeling. IEEE Transactions on Information Theory, 44(6):2743-2760, 1998.

Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19:3-11, 1966.

Baumgartner TA. Estimating reliability when all test trials are administered on the same day. Research Quarterly, 40: 222-225, 1969.

Benjamini Y and Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of Royal Statistical Society Series B, 57: 289-300, 1995.

Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of measurements. Lancet, i: 307-310, 1986.

Bland JM and Altman DG. Statistics notes: Measurement error. British Medical Journal (BMJ), 313:744, 1996a.

Bland JM and Altman DG. Statistics notes: Measurement error and correlation coefficients. British Medical Journal (BMJ), 313:41-42, 1996b.

Brewer M. Research Design and Issues of Validity. In Reis H and Judd C (eds.): Handbook of Research Methods in Social and Personality Psychology. Cambridge University Press, Cambridge, UK, 2000.

British standard institution. Precision of test methods I: Guide for the determination and reproducibility for a standard test method. British Standard 5497, 1979.

Brown W. Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3: 296-322, 1910.

Carroll RJ, Ruppert D, Stefanski LA. Measurement Error in Nonlinear Models. Chapman & Hall, London, UK, 1995.

Carson RE. Precision and accuracy considerations of physiological quantitation in PET. Journal of Cerebral Blood Flow and Metabolism, 11: A45-A50, 1991.

Chan GL, Holden JE, Stoessl AJ, Doudet DJ, Wang Y, Dobko T, Morrison KS, Huser JM, English C, Legg B, Schulzer M, Calne DB, Ruth TJ. Reproducibility of the distribution of carbon-11-SCH 23390, a dopamine D1 receptor tracer, in normal subjects. The Journal of Nuclear Medicine 39: 792-797, 1998.

Charter RA. Effect of measurement error on tests of statistical significance. Journal of clinical and experimental neuropsychology, 19: 458-462, 1997.

Cherry SR. Fundamentals of positron emission tomography and applications in preclinical drug development. Journal of Clinical Pharmacology, 41: 482-491, 2001.

Chow TW, Takeshita S, Honjo K, Pataky CE, St Jacques PL, Kusano ML, Caldwell CB, Ramirez J, Black S, Vehoeff NPLG. Comparison of manual and semi-automated delineation of regions of interest for radioligand PET imaging analysis. BMC Nuclear Medicine, 7:2, doi: 10.1186/1471-2385-7-2, 2007.

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20:37-46, 1960.

Cohen J. Statistical Power Analysis for the Behavioral Sciences. Academic Press, New York, USA, First edition, 1969.

Cohen J. The earth is round ($p<.05$). American Psychologist, 49(12):997-1003, 1994.

Cox DR and Hinkley DV. Theoretical Statistics. Chapman and Hall, London, UK, First edition, 1974.

Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika,16:407-424, 1951.

Cselényi Z, Olsson H, Halldin C, Gulyás B, Farde L. A comparison of recent parametric neuroreceptor mapping approaches based on measurements with the high affinity PET radioligands [^{11}C]FLB 457 and [^{11}C]WAY 100635. NeuroImage, 32:1690-1708, 2006.

Cselényi Z, Olsson H, Farde L, Gulyás B. Wavelet-aided parametric mapping of cerebral dopamine D_2 receptors using the high affinity PET radioligand [^{11}C]FLB 457. NeuroImage, 17(1):47-60, 2002.

DerSimonian R and Kacker R. Random-effects model for meta-analysis of clinical trials: An update. Contemporary Clinical Trials, 28:105-114, 2007.

Desmond JE, Glover GH. Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. Journal of Neuroscience Methods, 118: 115-128, 2002.

D'Esposito M. Functional neuroimaging of cognition. Seminars in Neurology, 20(4): 487-498, 2000.

Eckelman WC. Accelerating drug discovery and development through in vivo imaging. Nuclear Medicine and Biology, 29: 777-782, 2002.

Erdfelder E, Faul F, Buchner A. GPOWER: a general power analysis program. Behavior Research Methods, Instruments & Computers, 28(1): 1-11, 1996.

Fessler JA. Penalized weighted least squares least-squares image reconstruction for positron emission tomography. IEEE Transactions on Medical Imaging, 13: 290-300, 1994.

Fisher RA. The Design of Experiments. Oliver and Boyd, Edinburgh, UK, First edition, 1934.

Fisher RA. Statistical Methods and Scientific Inference. Macmillan Publishing, New York, USA, Third edition, 1973.

Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76:378-382, 1971.

Fleiss JL. Analysis of data from multiclinic trials. Controlled clinical trials, 7: 267-274, 1986.

Friston KJ, Holmes AP, Poline J-B, Price CJ, Frith CD. Detecting activations in PET and fMRI: levels of inference and power. NeuroImage, 4:223-235, 1996.

Friston KJ, Holmes AP, Worsley KJ, Poline J-B, Frith CD, Frackowiak RSJ. Statistical parametric maps in functional imaging: a general linear approach. Human Brain Mapping, 2: 189-210, 1995.

Friston KJ, Holmes AP, Worsley KJ. How many subjects constitute a study? Neuroimage, 10(1):1-5, 1999.

Friston KJ, Penny WD, Phillips C, Kiebel SJ, Hinton G, Ashburner J. Classical and Bayesian inference in functional neuroimaging: Theory. NeuroImage, 16: 465-483, 2002.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, Evans AC. Assessing the significance of focal activations using their spatial extent. Human Brain Mapping, 1: 214-220, 1994.

Genovese CR, Lazar NA, Nichols TE. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage, 15: 870-878, 2002.

Good IJ. What are degrees of freedom? The American Statistician, 27(5):227-228, 1973.

Grabowski TJ, Frank RJ, Brown CK, Damasio H, Boles Ponto LL, Watkins GL, Hichwa RD. Reliability of PET activation across statistical methods, subject groups and sample sizes. Human Brain Mapping, 4(1):23-46, 1996.

Griffin JI. Statistics: Methods and Applications. Holt, Rinehart and Winston, New York, USA, 1962.

Gunn RN, Lammertsma AA, Hume SP, Cunningham VJ. Parametric imaging of ligand-receptor binding in PET using a simplified reference tissue model. NeuroImage, 6:279-287, 1997.

Hacking I. Logic of Statistical Inference. Cambridge University Press, London, UK, 1965.

Hampel FR, Ronchetti EM, Rousseeuuw PJ, Stahel WA. Robust Statistics: The Approach Based on Influence Functions. John Wiley & Sons, New York, USA, 1986.

Hanley JA and McNeil BJ. The meaning and use of the area under receiver operating characteristic (ROC) curve. Radiology, 143:29-36, 1982.

Happonen AP and Alenius S. A comparison of sinogram and stackgram domain filtering methods employing L-filters for noise reduction of tomographic data. Proceedings of the 2005 Finnish Signal Processing Symposium, pp. 1-4, 2005.

Hedges LV and Pigott TD. The power of statistical tests in meta-analysis. Psychological Methods, 6(3):203-217, 2001.

Hochberg Y and Tamhane AC. Multiple Comparison Procedures. John Wiley & Sons, New York, USA, 1987.

Hirvonen J, Aalto S, Lumme V, Någren K, Kajander J, Vilkman H, Hagelberg N, Oikonen V, Hietala J. Measurement of striatal and thalamic dopamine D_2 receptor binding with ^{11}C-raclopride. Nuclear Medicine Communications, 24:1207-1214, 2003.

Hoffman EJ, Huang SC, Phelps ME. Quantitation in positron emission computed tomography: 1. Effect of object size. Journal of Computer Assisted Tomography, 3: 299-308, 1979.

Holmes AP. Statistical issues in functional brain mapping, PhD thesis, Department of Statistics, University of Glasgow, UK, 1994.

Holmes AP, Blair RC, Watson JDG, Ford I. Nonparametric analysis of statistic images from functional mapping experiments. Journal of Cerebral Blood Flow and Metabolism, 16(1):7-22, 1996.

Holmes AP and Friston KJ. Generalizability, random effects and population inference. NeuroImage 7, S754, 1998.

Hopkins, WG. Measures of reliability in sports medicine and science. Sports Medicine 30(1): 1-15, 2000.

Hripcsak G and Heitjan DF. Measuring agreement in medical informatics reliability studies. Journal of Biomedical Informatics, 35: 99-110, 2002.

Huesman RH. A new fast algorithm for the evaluation of regions of interest and statistical uncertainty in computed tomography. Physics in Medicine and Biology, 29: 543-552, 1984.

Hunt RJ. Percent agreement, Pearson's correlation, and Kappa as measures of inter-examiner reliability. Journal of Dental Research, 65(2):128-130, 1986.

International Organization for Standardization. Statistics - Vocabulary and Symbols. Part 1: Probability and general statistical terms. ISO 3534-1. Geneva: ISO, 1993.

Jackson DA, Somers KM, Harvey HH. Similarity coefficients: Measures of co-occurrence and association or simply measures of occurrence. The American Naturalist, 133(3):436-453, 1989.

Jain A. Fundamentals of Digital Image Processing. Prentice-Hall International, Englewood Cliffs, NJ, USA, 1989.

Jernigan TL, Gamst AC, Fennema-Notestine C, Ostergaard AL. More "mapping" in brain mapping. Human Brain Mapping, 19(2):90-95, 2003.

Kao C-M, Wernick MN, Chen C-T. Kalman sinogram restoration for fast and accurate PET image reconstruction. IEEE Transactions on Nuclear Science, 45(6,Part2):3022-3029, 1998.

Kay SM. Fundamentals of Statistical Signal Processing. Volume II: Detection theory. Prentice Hall PTR, New Jersey, USA, 1998.

Kelder SH, Jacobs DR, Jeffery RW, McGovern PG, Forster JL. The worksite component of variance: design effects and the Healthy Worker Project. Health Education Research, 8(4): 555-566, 1993.

Kim JS, Ichise M, Sangare J, Innis RB. PET imaging of serotonin transporters with [^{11}C]DASB: Test-retest reproducibility using a multilinear reference tissue parametric imaging method. The Journal of Nuclear Medicine, 47(2):208-214, 2006.

Kraemer HC and Robinson TN. Are certain multicenter randomized clinical trial structures misleading clinical and policy decisions? Contemporary Clinical Trials 26: 518-529, 2005.

Kraemer HC and Thiemann S. How Many Subjects? Statistical Power Analysis in Research. Sage, Newbury Park, USA, 1987.

Krantz DH. The null hypothesis controversy in psychology. Journal of the American Statistical Association, 44(448):1372-1381, 1999.

Kuder GF and Richardson MW. The theory of the estimation of test reliability. Psychometrika, 2: 151-160, 1937.

LaRiviére PJ and Pan X. Nonparametric regression sinogram smoothing using a roughness-penalized Poisson likelihood objective function. IEEE Transactions on Medical Imaging, 19(8):773-786, 2000.

Laruelle M. Modelling: when and why? European Journal of Nuclear Medicine, 26: 571-572, 1999.

Laruelle M. Imaging synaptic neurotransmission with in vivo binding competition techniques: A critical review. Journal of Cerebral Blood Flow and Metabolism, 20:423-451, 2000.

Lee C-M and Farde L. Using positron emission tomography to facilitate CNS drug development. Trends in Pharmacological Sciences, 27(6): 310-316, 2006.

Lee PM. Bayesian Statistics: An Introduction. Wiley, New York, USA, 1997.

Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? Journal of the American Statistical Association, 88(424):1242-1249, 1993.

Lindgren BW. Statistical Theory. The Macmillan Company, Toronto, Canada, Second edition, 1968.

Logan J, Fowler JS, Volkow ND, Wang G-J, Ding Y-S, Alexoff DL. Distribution volume ratios without blood sampling from graphical analysis of PET data. Journal of Cerebral Blood Flow and Metabolism, 16:834-840, 1996.

Mantel N and Haenszel WH. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22: 719-748,1959.

Mayo DG and Cox DR. Frequentist statistics as a theory of inductive inference. IMS Lecture Notes--Monograph Series, 49:77-97, 2006.

Meinert CL. Organization of multicentre clinical trials. Controlled Clinical Trials, 1:305-312,1981.

Mesina CT, Boellaard R, Jongbloed G, van der Vaart AW, Lammertsma AA. Experimental evaluation of iterative reconstruction versus filtered back projection for 3D [^{15}O]water PET activation studies using statistical parametric mapping analysis. NeuroImage, 19:1170-1179, 2003.

Neyman J and Pearson ES. On the problem of most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London, Ser. A, 231:289-337, 1933.

Nichols TE and Hayasaka S. Controlling the familywise error rate in functional neuroimaging: a comparative review. Statistical Methods in Medical Research, 12:419-446, 2003.

Nichols TE and Holmes AP. Nonparametric permutation tests for functional neuroimaging: A primer with examples. Human Brain Mapping, 15:1-25, 2002.

Ollinger JM and Fessler JA. Positron-emission tomography. IEEE Signal Processing Magazine, 14(1):43-55, 1997.

O'Rourke K. An historical perspective on meta-analysis: dealing quantitatively with varying study results. The James Lind Library (www.jameslindlibrary.org), 2006. Accessed 17 April 2007.

Paans AMJ, van Waarde A, Elsinga PH, Willemsen ATM, Vaalburg W. Positron emission tomography: the conceptual idea using a multidisciplinary approach. Methods, 27:195-207, 2002.

Pajevic S, Daube-Witherspoon ME, Bacharach SL, Carson RE. Noise characteristics of 3-D and 2-D PET images. IEEE Transactions on Medical Imaging, 17(1):9-23, 1998.

Parsey RV, Slifstein M, Hwang DR, Abi-Dargham A, Simpson N, Mawlawi O, Guo N-N, Van Heertum R, Mann JJ, Laruelle M. Validation and reproducibility of measurement of 5-HT_{1A} receptor parameters with [carbonyl-^{11}C]WAY-100635 in humans : comparison of arterial and reference tissue input functions. Journal of Cerebral Blood Flow and Metabolism, 20: 1111-1133, 2000.

Passchier J, Gee A, Willemsen A, Vaalburg W, van Waarde A. Measuring drug-related receptor occupancy with positron emission tomography. Methods, 27:278-286, 2002.

Pearson K. Mathematical contributions to the theory of evolution – III. Regression, heredity and panmixia. Philosophical Transactions of the Royal Society of London, Series A, 187:253-318, 1896.

Peltonen S and Ruotsalainen U. New sinogram filter design utilizing sinusoidal trajectories. Proceedings of IEEE Medical Imaging Conference (MIC2006), pp. 2770-2774, 2006.

Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, New York, USA, 2003.

Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: The impact of measurement error on sample size requirements in clinical trials. Biological Psychiatry, 47: 762-766, 2000.

Petersson KM, Nichols TE, Poline JB, Holmes AP. Statistical limitations in functional neuroimaging, II. Signal detection and statistical inference. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences, 354:1261-1281, 1999.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. British Journal of Cancer, 34:585-612, 1976.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. British Journal of Cancer, 35:1-39, 1977.

Phelps ME. Positron emission tomography provides molecular imaging of biological processes. Proceedings of the National Academy of Sciences of United States of America, 97:9266-9233, 2000.

Pocock SJ. Clinical Trials: A Practical Approach. John Wiley & Sons, 11th reprint, Chichester, UK, 1993.

Rae G. The equivalence of multiple rater kappa statistics and intraclass correlation coefficients. Educational and Psychological Measurement, 48: 367-374, 1988.

Reilhac A, Lartizien C, Costes N, Sans S, Comtat C, Gunn R, Evans AC..PET-SORTEO: A Monte Carlo based simulator with high count rate capabilities. IEEE Transactions on Nuclear Science 51(1): 46-52, 2004.

Reinders AATS. From methods to meaning in functional neuroimaging, PhD thesis, Medical Faculty, Neurology, University of Groningen, The Netherlands, 2004.

Reinders AATS, Paans AMJ, de Jong BM, den Boer JA, Willemsen ATM. Iterative versus filtered backprojection reconstruction for statistical parametric mapping of PET activation measurements: A comparative case study. NeuroImage, 15: 175-181, 2002.

Rencher AC. Linear Models in Statistics. John Wiley & Sons, Inc., New York, USA, 2000.

Rissanen J. Modeling by the shortest data description. Automatica, 14:465-471, 1978

Rorden C, Bonilha L, Nichols TE. Rank-order versus mean based statistics for neuroimaging. Neuroimage, 35: 1531-1537, 2007.

Rousset OG, MA Y, Evans AC. Correction for partial volume in PET: principle and validation. Journal of Nuclear Medicine, 39:904-911, 1998.

Rousson V, Gasser T, Seifert B. Assessing interrater and test-retest reliability of continuous measurements. Statistics in Medicine, 21:3431-3446, 2002.

Royall, R. Statistical Evidence: A Likelihood Paradigm. Chapman & Hall, London, 1997.

Ruotsalainen U, Mykkänen J, Luoma J, Tohka J, Alenius S. Methods to improve repeatability in quantification of brain PET images. World Congress on Neuroinformatics, Congress Proceedings ARGESIM Report no.20, ARGESIM/ASIM Verlag, Vienna. pp. 659-664, 2001.

Rusjan P, Mamo D, Ginovart N, Hussey D, Vitcu I, Yasuno F, Tetsuya S, Houle S, Kapur S. An automated method for the extraction of regional data from PET images. Psychiatry Research: Neuroimaging, 147: 79-89, 2006.

Saad ZS, Chen G, Reynolds RC, Christidis PP, Hammett KR, Bellgowan PSF, Cox RW. Functional Imaging Analysis Contest (FIAC) analysis according to AFNI and SUMA. Human Brain Mapping, 27: 417-424, 2006.

Saleem A, Charnley N, Price P. Clinical molecular imaging with positron emission tomography. European Journal of Cancer, 42: 1720-1727, 2006.

Scheffé H. The Analysis of Variance. Wiley and Sons New York, USA, pp. 221-260, 1959.

Shadish W, Cook T, Campbell D. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston, USA, 2002.

Shrier I, Platt RW, Steel RJ. Mega-trials vs. meta-analysis: precision vs. heterogeneity. Contemporary Clinical Trials, 28: 324-328, 2007.

Shrout PE and Fleiss JL. Intraclass correlation: uses in assessing rater reliability. Psychological Bulletin, 36: 420-428, 1979.

Shrout PE. Measurement reliability and agreement in psychiatry. Statistical Methods in Medical Research, 7:301-317, 1998.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage, 23(S1):S208-S219, 2004.

Snedecor GW and Cochran WG. Statistical Methods. The Iowa State University Press, Ames, USA, Sixth Edition, 1967.

Solomonoff RJ. A formal theory of inductive inference, part I. Information and control, 7(1):1-22, 1964a.

Solomonoff RJ. A formal theory of inductive inference, part II. Information and control, 7(2):224-254, 1964b.

Spearman C. Correlation calculated from faulty data. British Journal of Psychology, 3: 271-295, 1910.

Stevens J. Applied Multivariate Statistics for the Social Sciences. Lawrence Erlbaum Associates, Publishers, New Jersey, USA, 1986.

Tarkkonen L and Vehkalahti K. Measurement errors in multivariate measurement scales. Journal of Multivariate Analysis, 96:172-189, 2005.

Tauscher J, Verhoeff, N, Christensen BK, Hussey D, Meyer JH, Kecojevic A, Javanmard M, Kasper S, Kapur S. Serotonin 5-HT1A receptor binding potential declines with age as measured by [11C]WAY-100635 and PET. Neuropsychopharmacology 24:522-530, 2001.

Tohka J. Global optimization-based deformable meshes for surface extraction from medical images, PhD thesis, Institute of Signal Processing, Tampere University of Technology, Finland, 2003.

Trochim WM. The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: <http://www.socialresearchmethods.net/kb/> (version current as of October 20, 2006).

Turkheimer FE, Aston JAD, Cunningham VJ. On the logic of hypothesis testing in functional imaging. European Journal of Nuclear Medicine and Molecular Imaging, 31:725-732, 2004.

Turkheimer FE, Brett M, Aston JAD, Leff AP, Sargent PA, Wise RJS, Grasby PM, Cunningham VJ. Statistical modeling of PET images in wavelet space. Journal of Cerebral Blood Flow and Metabolism, 20:1610-1618, 2000.

Turkheimer FE, Smith CB, Schmidt K. Estimation of the number of "true" null hypotheses in multivariate analysis of neuroimaging data. Neuroimage, 13:920-930, 2001.

Turkington TG. Introduction to PET instrumentation. Journal of Nuclear Medicine Technology, 29: 1-8, 2001.

Van Horn JD, Ellmore TM, Esposito G, Berman KF. Mapping voxel-based statistical power on parametric images. NeuroImage, 7: 97-107, 1998.

Wahl LM and Nahmias C. Statistical power analysis for PET studies in humans. The Journal of Nuclear Medicine, 39(10):1826-1829, 1998.

Wallace CS and Boulton DM. An information measure for classification. Computer Journal, 11: 185-195, 1968.

Wallace CS and Boulton DM. Estimation and inference by compact coding. Journal of Royal Statistical Society, Series B, 49:240-265, 1987.

Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. Journal of Strength and Conditioning Research, 19(1): 231-240, 2005.

Werts CE, Rock RD, Linn RL, Jöreskog KG. A general method of estimating the reliability of a composite. Educational and Psychological Measurement, 38(4):933-938, 1978.

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC. A unified statistical approach for determining significant signals in images of cerebral activation. Human Brain Mapping, 4:58-73, 1996.

Yasuno F, Sanabria SM, Burns D, Hargreaves RJ, Ghose S, Ichise M, Chin FT, Morse CL, Pike VW, Innis RB. PET imaging of neurokinin-1 receptors with [^{18}F]SPA-RQ in human subjects: Assessment of reference tissue models and their test-retest reproducibility. Synapse 61: 242-251, 2007.

Yavuz M and Fessler JA. Statistical image reconstruction methods for randoms-precorrected PET scans. Medical Image Analysis, 2: 369-378, 1998.

Zijdenbos A, Forghani R, Evans A. Automatic quantification of MS lesions in 3D MRI brain data sets: Validation of INSECT. In Proc. of Medical Image Computing and

Computer-Assisted Intervention (MICCAI98), Lecture Notes in Computer Science 1496: 439-448, 1998.

Publications

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O. Box 527
FIN-33101 Tampere, Finland