Data-driven intensity normalization of PET group comparison studies is superior to global mean normalization.

Per Borghammer¹, Joel Aanerud¹, and Albert Gjedde¹,².

¹PET Center, Aarhus University Hospitals, Denmark
²Center of Functionally Integrative Neuroscience (CFIN), Aarhus University, Denmark

February 13th, 2009

Corresponding author
Per Borghammer, M.D., Ph.D.
PET Centre, Aarhus University Hospitals
Aarhus C, Denmark 8000
Email: per@pet.auh.dk
Phone: +0045 8949 4378
Fax: +0045 8949 4400

Short Title: Data-driven intensity normalization in PET group comparisons

Key Words: Parkinson's disease, energy metabolism, CBF, normalization, proportional scaling
ABSTRACT

Background: Global mean (GM) normalization is one of the most commonly used method of normalization in PET and SPECT group comparison studies of neurodegenerative disorders. It requires that no between-group GM difference is present, which may be strongly violated in neurodegenerative disorders. Importantly, such GM differences often elude detection due to the large intrinsic variance in absolute values of cerebral blood flow or glucose consumption. Alternative methods of normalization are needed for this type of data.

Materials & Methods: Two types of simulation were performed using CBF images from 49 controls. Two homogeneous groups of 20 subjects were sampled repeatedly. In one group, cortical CBF was artificially decreased moderately (simulation I) or slightly (simulation II). The other group served as controls. Ratio normalization was performed using five reference regions: (1) Global mean; (2) An unbiased VOI; (3) Data-driven region extraction (Andersson); (4-5) Reference cluster methods (Yakushev et al.) Using voxel-based statistics, it was determined how much of the original signal was detected following each type of normalization.

Results: For both simulations, global mean normalization performed poorly, with only a few percent of the original signal recovered. Global mean normalization moreover created artificial increases. In contrast, the data-driven reference cluster method detected 65-95% of the original signal.

Conclusion: In the present simulation, the reference cluster method was superior to GM normalization. We conclude that the reference cluster method will likely yield more accurate results in the study of patients with early to moderate stage neurodegenerative disorders.
INTRODUCTION:

Positron emission tomography (PET) has been used to investigate the cerebral blood flow (CBF) and cerebral metabolic rate of glucose (CMRglc) in a range of brain disorders. However, the absolute values of CBF and CMRglc exhibit large intra- and inter-individual variation. The coefficient of variance (COV; SD/mean) is most often in the order of 15% in healthy elderly subjects (Leenders et al., 1990), but can be as high as 30% in neurodegenerative disorders such as Parkinson’s disease (PD) (Huang et al., 2007) and Alzheimer’s disease (AD) (Fukuyama et al., 1994). For this reason, ratio normalization of regional tracer uptake to a reference region has been a standard data preprocessing step for more than two decades, most commonly using the global mean as the reference value (Fox et al., 1988). When using this approach, signals of a much smaller magnitude can be detected using the modest sample sizes typical of PET studies, and it also obviates the necessity for invasive blood sampling.

Global mean normalization has the fundamental requirement that no between-group differences exist in the global mean. This prerequisite generally is violated in studies of neurodegenerative disorders (Eidelberg et al., 1990; Minoshima et al., 1995a). Importantly, a difference in global mean can be well below detection threshold in statistical group comparisons. Indeed, to reliably detect a global mean decrease of 10% in PET data ($\alpha=0.05$, power=0.90, COV 15-30%, two-sided test), sample sizes of 50-200 subjects per group are needed. However, an undetected difference of this magnitude robustly introduces bias into an analysis employing global mean normalization, with the subsequent creation of artificial hypermetabolism in conserved regions (Borghammer et al., 2008a). To avoid the potential bias of global mean normalization, alternative reference regions have been proposed, most
commonly normalization to an \textit{a priori} defined brain region spared by the disease processes. Thus, the pons and cerebellum have been widely used for normalization of PET and single photon emission computed tomography (SPECT) studies of AD (Minoshima et al., 1995b; Soonawala et al., 2002) and PD (Pizzolato et al., 1988). In contrast to \textit{a priori} defined regions, Andersson proposed a data-driven method to define the normalization reference region \textit{a posteriori} (Andersson, 1997). And recently, Yakushev and colleagues reported a simple and elegant way to define a conserved region suitable as normalization reference region in comparisons of patients with mild AD and controls (Yakushev et al., 2009). This method provided very high accuracy in discriminating mild AD patients from normal controls, and has general applicability to comparisons of patients with neurodegenerative disorders with normal controls (see methods section for description of the methods).

Several comparative studies of the performance of different normalization methods were published (for reviews, see (Arndt et al., 1996; Gullion et al., 1996)). Most of these studies utilized \textit{real data} from activation tasks, such as finger tapping (Strother et al., 1995), memory tasks (Arndt et al., 1996; McIntosh et al., 1996), or visual stimulation (Andersson, 1997). Using real data, however, is problematic, since the extent and magnitude of the true signal is \textit{de facto} unknown. Moreover, it is unclear, whether results obtained from these activation studies, characterized by spatially restricted, but large changes, are directly applicable to the case of neurodegenerative disorders that often are characterized by more widespread decreases - often of a smaller peak magnitude than in activation studies. Indeed, in a previously published study, we simulated isolated cortical decreases of 11\% in most of the neocortex with interspersed smaller clusters of 23\% decrease (Borghammer et al., 2008a). This
Data-driven intensity normalization in PET group comparisons

manipulation decreased the global mean value only by 8-10%, which was below detection threshold in a 20 “patients” versus 20 controls comparison. Global mean normalization of these simulation experiments resulted in the detection of few voxels with 11% decrease, and only some of the 23% decreased voxels were identified. We suggest that this simulation is similar to studies of early-moderate stage patients with neurodegenerative disorders, in which the regional decreases often are too small for the global mean decrease to be detected by standard statistical tests. Thus, global mean normalization have often been employed in studies of early-stage AD (Kawachi et al., 2006; Perneczky et al., 2007) and PD (Eidelberg et al., 1994; Ghaemi et al., 2002; Huang et al., 2007; Imon et al., 1999), and many of these studies reported no or only slight decreases – as was the case in our simulation studies.

The objective of the present paper was to conduct a formal comparison of different types of normalization. Specifically, we investigated the case in which the signal consisted of modest decreases in spatially widespread regions. For this purpose, we employed simulated data to control the magnitude and extent of the true signal with absolute certainty. In all analyses, we used ratio normalization (Fox et al., 1988), but the reference region utilized for the normalization procedure differed. In short, we normalized to (1) the global mean, (2) the mean of an unbiased volume of interest (VOI) defined a priori, and (3) the mean of reference regions extracted by two different data-driven methods (Andersson, 1997; Yakushev et al., 2009). The normalized data were then analyzed by standard voxel-based statistical procedures and it was determined how well each type of normalization facilitated the detection of the true signal present in the simulated data.
MATERIALS AND METHODS:
The full details of subject recruitment, MR and PET scanning protocols, and data preprocessing were published previously (Borghammer et al., 2008a) and only a brief account is given here.

Subjects
We utilized PET CBF scans from 49 healthy volunteers (32 male/17 female; age 34-72 y), all of whom had participated in previous protocols. Written informed consent had been obtained from all study subjects. The studies had been approved by the official science ethics committee, and were in accordance with the declaration of Helsinki.

Scanning procedures
MRI. A high resolution T1-weighted MR was acquired for most subjects with a 3.0 T Signa Excite GE Magnet using a 3DIR-fSPGR sequence (256x256, TE1=min full, TI=450, slice thickness=1.5 mm.) A few subjects were scanned with GE MR 1.5-T Echo Speed tomograph (3D-SPGR, 256x256, 1 Splap, NEX: 1, slice thickness 1.5mm).

PET. Each subject underwent one dynamic 21-frame [$^{15}$O]H$_2$O emission recording with arterial blood sampling. Recordings were performed in a quiet room with subjects resting in a supine position with open eyes in a quiet, darkened room. The scans were acquired in 3D mode with the ECAT EXACT HR 47 (CTI/Siemens) whole-body tomograph. Images were reconstructed as 128 × 128 matrices of 2 × 2 mm voxels using filtered back-projection with a 0.5 cycles$^{-1}$ ramp filter, followed by
Gaussian filter, resulting in an isotropic resolution of 7 mm. Tissue attenuation scans were performed using a rotating $^{68}$Ge source.

**Data analysis**

Parametric maps of CBF were calculated using the single step, two-compartment, weighted-integration method (Ohta et al., 1996), and were co-registered, via the subjects’ individual MR images, to common stereotaxic space (Talairach and Tournoux, 1988), using a combination of linear and non-linear transformations (Brabner, 2006).

**Simulation I.**

From the pool of 49 CBF images in common space, we randomly sampled two groups of 20 subjects each, based on the following criteria: The groups should display no between-group differences in sex- and age-distribution, or in global CBF values on an unpaired t-test. If any of these t-tests yielded a p-value below 0.50, the groups were discarded and two new groups were sampled. This sampling technique ensured the creation of highly homogeneous groups.

The CBF maps of one of the groups were then systematically manipulated, while the other group was left unaltered. In short, we designed a specific image volume in standard space consisting of voxels with three possible values: 1, 0.89, or 0.77 (Figure 1A). In this volume, a total of 27.4% of all intra-cerebral voxels were assigned the value 0.89, 12.4% had the value 0.77, while the remaining voxels had the value 1. Importantly, all voxels with values 0.89 or 0.77 were situated in the cerebral cortex. By multiplying each CBF map (in standard space) with this volume, we
artificially decreased most of the cortical voxels of the CBF map to either 89% or 77% of original values, while leaving the remaining cortical and all subcortical voxels unchanged. The magnitudes of the decreases were decided by pilot trials. The specific aim was to introduce cortical decreases, which would not produce detectable decreases in the global CBF (due to the large variation in quantitative data). The entire sampling procedure and subsequent manipulation was repeated four times to produce a total of four sets of two groups. The full details of the simulations are given in (Borghammer et al., 2008a).

**Simulation II.**

While simulation I was meant to emulate a heterogeneous, moderate involvement of the cortex, we performed an additional simulation to investigate the case, in which a weaker (and homogeneous) signal was present. We produced a second image volume almost identical to the first one, except that the voxels, which in simulation I had a value of 0.77, were now also assigned the value of 0.89 (Figure 2A). This produced a volume with a homogeneous 11% decrease in large cortical regions. All subsequent steps of simulation II was identical to simulation I (see above). Simulation II was performed only twice. In general, the manipulations gave rise to only slightly visible alterations in the pattern of the raw CBF images (see Supplementary Figure 1 on the publishers website).

**Normalization.**

All trials of simulation I and II were all analyzed using five different types of normalization: (1) Global mean normalization was performed by dividing each voxel value by the mean of all intracerebral voxels (excluding extra-cerebral and ventricular
Data-driven intensity normalization in PET group comparisons

voxels) (Fox et al., 1988). The mean value was extracted by a whole-brain mask. (2) Ratio normalization to the mean of an unbiased reference region. In the present simulated data, we left the white matter (WM) unaltered, so we chose this region for unbiased \textit{a priori} normalization. The WM mask was conservative and included only voxels at some distance from GM voxels to exclude spill over effect from the GM. Details on the definition of the whole-brain and WM were provided previously (Borghammer et al., 2008b). (3) The data was analyzed with the data-driven method originally proposed by Andersson (Andersson, 1997). In brief, the method works iteratively. In the first iteration, a standard voxel-based statistical analysis with global mean normalization is performed. In the second iteration, a new normalization reference region is constructed by masking the output t-map from the first iteration, i.e. excluding all extreme voxels ($t < -2$ and $t > 2$). The resultant mask now includes all voxels with t-values close to zero, and is used to again normalize the original, non-normalized data. Another voxel-based statistical analysis is performed. The output t-map from the second iteration forms the basis for the normalization mask of the third iteration, and so on. For all analyses in the present study, we performed four such iterations accepting the fourth iteration as the final result. (4) We also analyzed the data with the recently proposed \textit{reference cluster method} (Yakushev et al., 2009). This method is very similar to the Andersson method, but involves only two iterations. First, a standard global mean normalized voxel-based analysis is performed. In the second iteration, a new normalization mask is likewise defined on the basis of the output t-map from the first iteration. However, the t-map is thresholded differently, i.e. only “hypermetabolic” voxels with t-values > 2 (p<0.05) are included. Normalization of the original non-normalized data with the new mask (created from the thresholded output t-map from the first iteration) is then carried out,
and subsequent voxel-based analysis is performed to produce the final results. The validity of this method has the fundamental requirement that the seemingly hypermetabolic region identified in the first iteration, is in fact a conserved region, in which no between-group changes are present. It is assumed that the “hypermetabolic” region has been artificially inflated by biased global mean normalization, due to isolated cortical decreases in one group. (5) Finally, it could be argued that a more conservative thresholding (i.e. \( p<0.001 \)) of the output t-map from the first Yakushev iteration would identify an even more conserved region, see (Yakushev et al., 2009). However, since a very stringent threshold would necessarily lead to identification of a much smaller reference region, this method could potentially be vulnerable to random noise in the data set, with resultant detection of too extensive regional decreases (i.e. false positive decreases, not to be confused with the false positive increases detected by biased global mean normalization). We investigated this potential problem by performing an additional Yakushev type normalization, in which the output t-maps were thresholded at \( t>3.6 \) (\( p<0.001 \)). The two Yakushev methods are referred to as Yakushev\( _2 \) and Yakushev\( _{3.6} \), respectively.

**Statistical analysis.**

The global values of all CBF images prior to, and subsequent to manipulations were extracted, and group comparisons were performed using unpaired t-tests.

Prior to voxel-based analysis, the coregistered, normalized CBF maps were blurred with a Gaussian filter to a resultant full-width-at-half-maximum (FWHM) of 14x14x14 mm. We analyzed the data with univariate statistics using the freely available software package *FMRISTAT* written by Worsley and colleagues (available
at www.math.mcgill.ca/keith/fmristat). This method is nearly identical to the commonly used SPM methodology. In the analysis, we used the mixed effect model analysis method as advocated in Section Four of Worsley et al. (Worsley et al., 2002), with spatial smoothing of the standard deviation image to increase the degrees of freedom. Appropriate linear contrasts were defined to reveal group differences in CBF. FMRISTAT assigns a t-value to each voxel in the brain and examines the map for significant focal changes (p < 0.05, corrected for false discovery rate (Genovese et al., 2002)), based on 3D Gaussian Random Field Theory (Worsley, 1996).

To estimate each normalization method’s ability to facilitate detection of the true signal, the number of voxels identified in each voxel-based analysis was compared to the actual number of truly decreased voxels present in the manipulation image volume. This is presented as a percentage in the results section, i.e. 100 x (number of voxels detected within the pattern of truly decreased voxels / total number of truly decreased voxels in the manipulation image volume). For simulation I, we separately identified the percentage of severely (77%) and moderately (89%) decreased voxels identified by the voxel-based analysis. When a normalization method gave rise to artificial increases, the number of detected voxels displaying increases is presented as a percentage of the number of detected voxels showing decreases.

RESULTS

Global differences. There were no age-, sex-, or global CBF differences between groups before the manipulation in any of the trials in simulations I and II (p > 0.5 in all t-tests). The global CBF values prior to, and after the manipulation are presented in Table 1. In simulation I, the four manipulated groups had their global CBF decreased
8-10% as a consequence of the manipulation, but this was below detection threshold when comparing to the control groups (p>0.10). In both trials of simulation II, the manipulation resulted in a 5% decrease of the gCBF, which was also below detection threshold when comparing to controls (p>0.31).

**Voxel-based analyses.** Figures 1 and 2 visually illustrate the extent of the significant clusters detected in simulation I and II by the different types of normalization. Table 2 summarizes the percentage of the true signal recovered by the five normalization methods. In short, for both simulations I and II the methods performed as follows (listed worst to best): Global mean normalization; Andersson normalization; ratio normalization to white matter; Yakushev2; Yakushev3.6. Significant artificial increases were detected only following global mean normalization (Table 3). Finally, although the Yakushev3.6 method detected most of the true signal in all simulations, it also had the propensity to detect false significant decreases, i.e. voxels, which had not been artificially decreased by the manipulation step. This is illustrated in Figures 1F and 2F (yellow arrows).

**DISCUSSION**

The present study yielded two important findings. (1) Standard global mean normalization performed very poorly and detected only a few percent of the voxels subjected to an 11% decrease. Moreover, only GM normalization consistently created artificial increases. (2) By contrast, the reference cluster method developed by Yakushev et al performed extraordinarily well. A central part of the present simulations was to induce cortical decreases, which only slightly decreased the global mean values, i.e. by 8-10% in simulation I, and by 5% in simulation II. Nevertheless,
the reference cluster method correctly identified 65-95% of the slightly decreased (11%) voxels, and more than 91% of the severely decreased (23%) voxels (Table 2).

In the following paragraphs the different normalization methods will be discussed.

**White matter and Andersson normalization.**

WM normalization can be considered a gold standard normalization procedure in the present simulation, since it was left unaltered by the manipulation. Indeed, the clusters detected after WM normalization were most often 300-500% larger then clusters detected subsequent to GM normalization (Table 2). Nevertheless, only 33% in average of the slightly decreased voxels (11%) were identified. In itself, this is not very informative since the absolute number of voxels detected is dependent on the power of the individual study, i.e. sample size and data variance. Nevertheless, it illustrates that in a typical CBF PET comparison of two groups of 20 subjects, gold standard ratio normalization to an unbiased region does not guarantee the reliable detection of widespread, low-magnitude decreases.

The data-driven method developed by Andersson (Andersson, 1997) generally performed better than GM normalization, but not quite as well as WM normalization for this kind of data. This was actually to be expected. The following idealized example illustrates how the iterative Andersson method can be trapped. Consider a group comparison in which one group displays heterogeneous decreases, i.e. one third of the brain is decreased by 20%, one third by 10%, while the remaining third is unchanged. The decreases are homogenous across tissue types, so the global mean is decreased by 10%. The first Andersson iteration, i.e. standard GM normalization yields a t-map, in which the unchanged region appears hypermetabolic (t>2), while
only the 20% decreased region will be identified as hypometabolic (t<-2). These regions are excluded in the second Andersson iteration, which retains only the apparently unchanged region (t-values close to zero). However, this region was in reality decreased by 10%, and the subsequent iterations will be identical to the first one. Thus, the Andersson method is trapped. In this idealized case, the performance of Andersson normalization would be identical to GM normalization. In the present study, it actually performed much better – almost equal to WM normalization. Indeed, the final normalization mask (Andersson iteration three) was somewhat similar to our a priori defined WM mask, although some overlap with the truly decreased regions were detected. All in all, Andersson normalization is surely preferable to standard GM normalization, and seems recommendable, when no a priori expectations of the data is available, which precludes the use of VOI normalization, or the Yakushev method. A comparison of the normalization masks employed in the WM, Andersson, and Yakushev is provided as supplementary material (Supplementary Figure 2).

It should be noted that we used only four iterations, whereas ten iterations were used in the original study by Andersson. However, there was hardly any difference in the final results between the third and fourth iteration in any of the six simulation trials, so we concluded that the “ceiling” was reached by the fourth iteration in our data sets.

**Applicability and caveats of the reference cluster method**

The Yakushev reference cluster method performed very well and has potential applicability in several cases: (a) group-comparisons of control subjects to patients, in whom isolated uni-directional changes exist (i.e. isolated decreases or increases), (b) intervention studies, in which the treatment induces regional uni-directional changes
in some brain regions while leaving other regions unchanged, (c) group-comparisons of control subjects to patients, in whom bi-directional changes exist (i.e. both decreases and increases), but only for cases in which one direction dominates. This is the case in PD, in which many human PET studies report absolute decreases in CBF and CMRglc in widespread cortical regions. Although no absolute subcortical increases were ever convincingly reported in the patient studies, such absolute subcortical increases were reported in autoradiography studies of animal models of PD - in isolated small basal ganglia structures (pallidum, thalamic subnuclei, pedunculo-pontine nucleus; see (Borghammer et al., 2008a) for references and full discussion). Thus, for application of the Yakushev method to PD, these subcortical structures should be excluded from the final normalization reference region.

Some caveats must be mentioned. The accuracy of the reference cluster method is sensitive to the specific t-threshold and the optimal t-threshold surely varies between different applications. In our simulation, the ground truth was known, so the optimal threshold could potentially be determined from an ROC-analysis. However, this is not the case for real data. A too stringent threshold leads to the definition of a small normalization region, which is inherently susceptible to random noise. In this scenario, the Yakushev method will be overly optimistic, i.e. lead to identification of false decreases. In the present simulation, the propensity to detect false decreases was only a major concern at the restricted threshold of t>3.6. Based on the present study, we recommend that a threshold of not more than t>2 (p<0.05 uncorrected) is used in most data sets, particularly in studies of early-stage disease, in which between-group GM differences are smallest. Further simulation studies are needed for a full description of the stability and performance of the Yakushev method, so care should
be taken in the interpretation of results based on this method. Yet, this particular
caveat applies to all normalization methods. Indeed, the present study confirms that
whereas the Yakushev method may lead to the detection of false positive decreases
(type I error), the gold standard WM normalization method lead to a marked
underestimation of the true pattern (type II error). In the study of neurodegenerative
disorders, investigators should therefore ideally utilize more than one type of
normalization, i.e. VOI- and Yakushev normalization, and interpret the results
accordingly. This also applies when automatic procedures are employed for
differential diagnostic purposes, i.e. the false positive decreases detected after
Yakushev normalization may create problems for correctly categorizing individual
scans in accordance with disease entities. However, in a recent study, the Yakushev
method exhibited the highest accuracy in correctly diagnosing patients with AD when
compared to VOI normalization (Yakushev et al., 2009), so it seems that the type I
error created by Yakushev normalization presents less of a problem than the type II
error created by unbiased VOI normalization.

**On absolute values and Parkinson’s disease.**

As mentioned in the introduction, samples sizes of 50-200 subjects pr group are
needed to detect a 10% decrease in the GM, and as many as 180-750 subjects per
group would be needed to detect a the 5% decrease present in simulation II ($\alpha=0.05,$
power=0.90, COV 15-30%). The typical PET or SPECT group comparison involves
sample sizes of 10-40 subjects pr group, making the detection of a 5-10% decrease in
the GM mean improbable. This is likely part of the explanation, why GM
normalization is one of the most commonly used methods. However, as mentioned
above, GM normalization could well be the worst possible method of normalization
for this type of data, whereas VOI methods and particularly the reference cluster method would likely yield more correct results. These considerations have the very serious implication that a large number of studies of neurodegenerative disorders may have yielded incomplete results, i.e. the extent of true hypometabolism has been underestimated. Moreover, artefactual hypermetabolism has been reported in regions, which were merely conserved.

In support of this conclusion, the following examples from the Parkinson’s disease (PD) literature will serve as an illustration. Global mean CBF and CMRglc values were reported in at least 23 comparisons of PD and healthy controls. Of these comparisons, seven reported significant GM decreases in PD patients (Bohnen et al., 2007; Globus et al., 1985; Hu et al., 2000; Imon et al., 1999; Karbe et al., 1992; Kuhl et al., 1984; Sasaki et al., 1992), thirteen reported non-significant decreases in PD (Abe et al., 2003; Agniel et al., 1991; Berding et al., 2001; Bes et al., 1983; Eidelberg et al., 1994; Ghaemi et al., 2002; Huang et al., 2007; Kitamura et al., 1988; Leenders et al., 1985; Montastruc et al., 1987; Otsuka et al., 1991; Perlmutter and Raichle, 1985; Playford et al., 1992), and only three studies reported small non-significant GM increases in the PD groups (Arahata et al., 1999; Huang et al., 2007; Otsuka et al., 1991). Moreover, nine additional studies disclosed absolute decreases in regional (mostly cortical) values, but did not explicitly report global mean values (Eberling et al., 1994; Eidelberg et al., 1990; Kondo et al., 1994; Mito et al., 2005; Otsuka et al., 1996; Peppard et al., 1992; Piert et al., 1996; Vander Borght et al., 1997; Wolfson et al., 1985). No study reported significant absolute increases, except one very early PET study of only four PD patients, which reported increases almost everywhere (Rougemont et al., 1984). Taken together, this evidence strongly suggests that cortical
absolute CBF and CMRglc is decreased in PD, and therefore by extension the global mean is also decreased – even if undetected in the individual study.

Nevertheless, most normalized PET and SPECT studies of PD employed GM normalization (Eidelberg et al., 1994; Huang et al., 2007; Imon et al., 1999; Nagano-Saito et al., 2004), and these reported extensive subcortical increases in the cerebellum, white matter, thalamus, pallidum, and striatum, with concomitant cortical decreases. In contrast, all 16 studies of PD, which utilized VOI normalization (to the cerebellum or pons), reported no subcortical increases, and quite extensive cortical decreases (see (Borghammer et al., 2008a) for references). This shift of pattern is predicted by the present simulation studies (compare Figure 1B and 1D), indicating that PD is characterized by widespread cortical decreases of activity with relative preservation of subcortical regions. Therefore, we predict that a reevaluation of populations of early-stage PD patients with the preferable normalization techniques presented in this study, will demonstrate more widespread cortical decreases in PD, and no concurrent widespread subcortical increases.

**Other types of normalization – ANCOVA & SSM.**

In this study, we only considered standard ratio normalization, which assumes a proportional relationship between regional and global values. However, a covariance adjustment using linear regression (ANCOVA) was developed for cognitive activation studies, upon demonstration that an additive model better approximates the relationship between regional and global values in cognitive activation PET studies (Friston et al., 1990). The fundamental requirement in ANCOVA normalization is that homogeneous regression coefficients exist between groups. However, covariance
adjustment with global mean as a covariate can reveal heterogeneous regression coefficients between groups of subjects (Devous et al., 1993; Gullion et al., 1996), which can be a serious limitation to the use of ANCOVA as an approach to removing intersubject variation in global mean values. Moreover, previous comparisons of ratio and ANCOVA normalization in cognitive activation studies (Arndt et al., 1996; McIntosh et al., 1996; Strother et al., 1995) reported that the two methods perform equally well. We previously reported that GM ratio normalization and ANCOVA normalization yield approximately the same results in studies of healthy aging (Borghammer et al., 2008b), i.e., they both create artificial increases in subcortical regions and detect smaller clusters of cortical decreases than VOI ratio normalization to white matter mean. Similarly, ANCOVA normalization of PD data (Eckert et al., 2005) leads to the detection of widespread subcortical increases, which is also seen after GM normalization, but never when using the cerebellum as the reference region. For these reasons, we chose not to employ ANCOVA normalization in the present study.

Statistical Parametric Mapping (SPM) (Friston, 1994) is the most commonly used voxel-based statistical approach, but some studies have been performed with alternative studies, such as the principal component network analysis known as the scaled subprofile model (SSM) (Moeller et al., 1987; Strother et al., 1995). Importantly, this method performs a preprocessing step that resembles GM normalization. Nevertheless, it has been claimed that SSM removes irrelevant global scaling factors, without introducing bias into the analysis (Ma et al., 2008; Strother et al., 1995). Yet, in our previous simulation studies we found the SSM to perform very similarly to standard GM normalization in SPM style analysis, i.e. artificial
subcortical increases were robustly created, whereas smaller clusters of cortical decreases were detected, than seen following unbiased VOI ratio normalization (Borghammer et al., 2008a; Borghammer et al., 2008b). Therefore, SSM was not considered further in the present study.

**Limitations of the simulation.**

Any simulation should be a reasonable approximation of reality, since generalization of the findings may otherwise be compromised. We do not claim that the present simulations accurately portray any specific neurodegenerative disorder. Nevertheless, we argue that the simulations were a realistic general approximation for the following reasons: (1) The simulated pattern of modest (11%) and severe (23%) cortical decreases were based on a t-map from an actual PD versus Controls comparison (see Borghammer et al., 2008a for details). (2) Previous studies of AD and PD display heterogeneous t-values, i.e. the classical regions, which are also the first to appear, consistently exhibit the highest t-values (Soonawala et al., 2002; Yakushev et al., 2009). At later disease stages, new regions appear but these display lower t-values than the primary regions. Thus, simulation I is probably the most realistic emulation of neurodegenerative disorders, whereas simulation II may be less realistic. (3) Several previous studies of AD compared GM normalization to alternative VOI normalization. They unanimously reported that normalization to regions known to be spared in AD (pons, motor cortex, cerebellum) resulted in the detection of much more extensive, but fairly isolated cortical decreases than seen following GM normalization (Minoshima et al., 1995a; Soonawala et al., 2002; Yakushev et al., 2009). And even more widespread cortical decreases were detected using the cluster reference method in AD (Yakushev et al., 2009). (4) Many researches claim that GM is not affected at
early stages of neurodegenerative disorders (Ma et al., 2008). However, as reviewed above, most quantitative studies of PD reported absolute decreases in cortical and global CBF and CMRglc. Absolute increases were never convincingly reported anywhere in the brain. Moreover, it is counterintuitive to assume that the GM decreases reported at later disease stages are not preceded by smaller (undetected) decreases at earlier disease stages, especially when considering that the absolute increases, which would be necessary to balance the often reported absolute decreases, have themselves never been reported. (5) Finally, it is impossible to know whether our simulation was a realistic emulation of the true extent and magnitude of cortical decreases in early stage PD. However, some VOI based SPECT CBF studies employed cerebellum ratio normalization. Assuming that the cerebellum is fairly conserved in PD, the relative between-group differences in different regions would be similar to the true absolute differences. One study compared 20 healthy controls to 17 early-stage PD patients without cognitive impairment. Relative decreases of 8-12% were detected in the four brain lobes (Derejko et al., 2006). This is comparable to the present simulation I.

We specifically chose not to alter the subcortical regions, since much evidence indicates that subcortical regions are relatively conserved in conditions such as Alzheimer’s disease (Buchert et al., 2005), Parkinson’s disease (Berding et al., 2001; Hu et al., 2000), healthy aging (Kalpouzos et al., 2007), and hepatic encephalopathy (Borghammer et al., 2008b). A few studies reported absolute thalamic and striatal decreases. However, many GM normalized studies detected relative increases in these and other subcortical regions, suggesting that any decrease in subcortical regions is of
a smaller magnitude than the decrease in the GM. For simplicity, we chose not to perturb the subcortical regions in our simulation.

Finally, we used CBF images from healthy controls, which have a slightly poorer signal/noise ratio than FDG-PET images. We also filtered our CBF images to a resultant FWHM of 14mm, whereas FDG-PET images often employ filters of 10-12 mm. However, the aim of our study was to investigate how different normalization methods affect the detection of large, widespread clusters of change in signal. Therefore, the results should be robust irrespective of filter size used, and should generalize to both FDG-PET (smaller filters) and SPECT-CBF studies (larger filters).

**Summary**

We repeatedly performed simulations of group-comparisons, in which one group had isolated cortical decreases. We contend this to be a realistic simulation of neurodegenerative disorders in general. Ratio normalization to five different reference regions were compared, and standard global mean normalization was found to perform very poorly in the detection of the true signal. Furthermore, it robustly created artificial increases in conserved regions. In contrast, the data-driven reference cluster method correctly identified most of the true signal without creating artificial increases. We conclude that many neurodegenerative disorders, such as Alzheimer’s disease, Parkinson’s disease, and other neurodegenerative disorders should be reevaluated using data-driven normalization methods. We predict that more widespread cortical decreases will be detected in these disorders, even at early disease-stages. This would have important implications for our understanding of the neuropathological mechanisms behind these disorders.
ACKNOWLEDGEMENTS

This work was supported by the Danish National Science Foundation, Medical Research Council of Denmark, and the Danish Parkinson Foundation. The authors wish to thank the reviewers for providing many helpful comments.

DISCLOSURE / CONFLICT OF INTEREST

None.

REFERENCES


Figure 1. Illustration of the five types of normalization in Simulation I (trial 2). A. The manipulation image volume used in simulation I. B. Global mean normalization produced large artefactual increases in all four trials and detected very little of the true signal. C. Andersson (AND) normalization. D. Ratio normalization to the mean of white matter (WM). E. Yakushev normalization using a liberal t>2 threshold (YAK$_2$). F. Yakushev normalization using a restricted t>3.6 threshold (YAK$_{3.6}$) detected most of the true signal, but also some “false significant decreases” (yellow arrows). [Note: the t-value scaling is extended in E-F, due to the very extreme t-values reported in the two Yakushev normalizations. All slices are $z = -1$ (MNI space).]
**Figure 2.** The five types of normalization in Simulation II (trial 2). **A.** The homogeneous manipulation image volume used in simulation II. **B-F.** Global mean and Andersson normalization identified very little of the true signal. VOI normalization identified slightly more. Both Yakushev methods recovered much more of the original signal, but the YAK$_{3.6}$ method identified even more false decreases (yellow arrows) than in simulation I. [See **Figure 1** for details and abbreviations. All slices are $z = 5$ (MNI space).]
Supplementary Figure 1. A. The two groups of 20 healthy subjects sampled in Simulation II. The large inter-individual variation (SD/mean = 20%) in mean absolute CBF values is typical for absolute CBF and cerebral rate of glucose (CMRglc) data in the literature. B. In one group (right) a large part of the cerebral cortex was artificially decreased by 11%. However, no striking between-group differences are discernible from the visual impression of the raw absolute CBF images after manipulation. [CBF units: mL/100g/min. Slices are visualized by the view_slices feature of fMRIstat.]
**Supplementary Figure 2.** The figure illustrates the true signal (dark blue and light blue) and the extent of the normalization masks (green) used for simulation I, trial 2. Top row depicts the Yakushev normalization mask (i.e. all voxels $t>2$ from the standard GM normalization analysis). Bottom row left illustrates the *a priori* defined white matter mask. Bottom row right illustrates the final Andersson mask (i.e. all voxels $-2<t<2$ from the third Andersson analysis).
Table 1. Global CBF values in the control and manipulated groups.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Controls</th>
<th>Manipulated Group</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>p-value</td>
</tr>
<tr>
<td></td>
<td>Before Manipulation</td>
<td>After Manipulation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
<td>p-value</td>
</tr>
<tr>
<td>Simulation I</td>
<td>1</td>
<td>37.5 6.8</td>
<td>37.4 6.8</td>
<td>0.95</td>
<td>34.2 6.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>40.0 9.3</td>
<td>38.9 7.4</td>
<td>0.69</td>
<td>35.7 6.8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>38.9 7.6</td>
<td>39.5 8.3</td>
<td>0.81</td>
<td>36.2 7.6</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>39.1 7.0</td>
<td>39.2 6.6</td>
<td>0.98</td>
<td>35.9 6.0</td>
</tr>
<tr>
<td>Simulation II</td>
<td>1</td>
<td>36.1 7.6</td>
<td>36.2 6.8</td>
<td>0.96</td>
<td>34.3 6.4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>37.1 8.6</td>
<td>36.3 7.8</td>
<td>0.76</td>
<td>34.5 7.4</td>
</tr>
</tbody>
</table>

Global mean CBF in the control groups was compared with that in the manipulated groups prior to, and after the manipulation (see methods). [CBF has units ml/100g/min. Unpaired two-sided t-tests were used.]
Table 2. The percentage of true signal recovered by the five normalization methods.

<table>
<thead>
<tr>
<th></th>
<th>SIMULATION I</th>
<th></th>
<th>SIMULATION II</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Global</td>
<td>77 24.9% 27.2% 25.1% 21.8%</td>
<td>24.7±2.2%</td>
<td>89 5.9% 6.7% 6.0% 4.6%</td>
<td>5.7±0.8%</td>
</tr>
<tr>
<td>AND</td>
<td>77 59.2% 56.7% 54.4% 47.5%</td>
<td>54.4±5.1%</td>
<td>89 20.9% 20.0% 19.6% 13.3%</td>
<td>18.4±3.4%</td>
</tr>
<tr>
<td>WM</td>
<td>77 73.6% 66.5% 88.2% 69.4%</td>
<td>74.4±9.6%</td>
<td>89 28.4% 24.8% 54.0% 24.6%</td>
<td>33.0±14.1%</td>
</tr>
<tr>
<td>YAK2</td>
<td>77 93.7% 95.8% 92.1% 91.5%</td>
<td>93.3±2.0%</td>
<td>89 74.8% 81.5% 71.0% 67.6%</td>
<td>73.7±6.0%</td>
</tr>
<tr>
<td>YAK3.6</td>
<td>77 95.5% 97.6% 96.0% 96.4%</td>
<td>96.4±0.9%</td>
<td>89 80.8% 87.4% 85.6% 82.9%</td>
<td>84.2±2.9%</td>
</tr>
</tbody>
</table>

For each normalization method, the two rows present the percentage of severe (77) and moderate (89) decreases detected. In simulation II a uniform decrease to 89% of original values were used. [WM=white matter. AND=Andersson normalization. YAK2=Yakushev normalization using liberal threshold of t>2. YAK3.6=Yakushev using threshold of t>3.6.]
Table 3. Artificial increases detected following global mean normalization.

<table>
<thead>
<tr>
<th></th>
<th>SIMULATION I</th>
<th></th>
<th>SIMULATION II</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Global</td>
<td>44.9%</td>
<td>29.9%</td>
<td>100.2%</td>
<td>38.9%</td>
</tr>
</tbody>
</table>

The extent of the artificial increases are presented as a percentage of the detected true decreases, i.e. 100 x (voxels showing increases / voxels showing decreases).