From THE DEPARTMENT OF CLINICAL NEUROSCIENCE
Karolinska Institutet, Stockholm, Sweden

RELIABILITY, REPLICABILITY AND REPRODUCIBILITY IN PET IMAGING

Granville J. Matheson

Karolinska Institutet

Stockholm 2018

Cover Illustration: Metaphorical representation of the concepts of reliability (left), replicability (centre) and reproducibility (right) depicted through cats. Created by Tasartir through fiverr. Explanations of the representations can be found in the Introduction section.

Published by Karolinska Institutet
Printed by US-AB
ISBN 978-91-7831-283-2

Reliability, Replicability and Reproducibility in PET Imaging

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Granville J. Matheson

Principal Supervisor:
Dr. Simon Cervenka
Karolinska Institutet
Department of Clinical Neuroscience

Co-supervisors:
Dr. Predrag Petrovic
Karolinska Institutet
Department of Clinical Neuroscience

Prof. Lars Farde
Karolinska Institutet
Department of Clinical Neuroscience

Opponent:
Prof. Robert Innis
National Institute of Mental Health
Molecular Imaging Branch

Examination Board:
Prof. Mark Lubberink
Uppsala Universitet
Department of Surgical Sciences

Dr. Jussi Hirvonen
University of Turku
Department of Radiology

Prof. Stefan Wiens
Stockholm University
Department of Psychology

Abstract

Positron Emission Tomography (PET) is a non-invasive biomedical imaging method which can quantify biochemical markers as well as functional and metabolic activity in vivo. Following image analysis, quantification can be performed using various pharmacokinetic models, or using simplified semi-quantitative methods. There are numerous methods by which PET data can be analysed, and outcomes which can be reported, which differ in their accuracy, stability and specificity. The quantification of PET data, as well as the statistical procedures used to test clinical hypotheses, leads to conclusions which may differ in their degree of correctness.

This thesis explores themes of reliability, replicability and reproducibility for PET research. Reliability concerns the consistency and accuracy of an outcome for distinguishing between individuals. Replicability concerns the accuracy of research conclusions, and whether they can be obtained again using the same procedures in new studies. Reproducibility concerns steps towards increasing the transparency of data analysis, by recording and sharing the exact procedures used to arrive at the conclusions. Across the studies in this thesis, these themes are detailed, expanded upon, and used in a series of methodological and applied clinical PET studies.

In **Study I**, the performance of surface-based methods for normalisation and smoothing of PET data were compared with volumetric methods for exploratory parametric analysis using PET test-retest data measuring [^{11}C]SCH23390 BP_{ND} in cortical regions. We replicated previous results of decreased spread, and showed that these methods also show improved test-retest repeatability. In **Study II**, using the same data we evaluated the performance of post-reconstruction movement correction as well as automatic and manual methods for delineation of regions of interest. We showed that motion correction improves the reliability and repeatability of binding estimates, and that automatic methods for delineation do not perform less well than manual methods, and appear to be more consistent.

Study III evaluated the test-retest performance of simplified ratio-based outcome measures for quantification of translocator protein (TSPO) binding using [^{11}C]PBR28. We showed that these methods exhibit poor reliability, and little to no association with the gold-standard outcome measure V_T, suggesting that caution is warranted for interpretation of studies making use of these measures.

In **Study IV**, diurnal and seasonal changes in the availability of the serotonin 1A receptor and the serotonin transporter were measured across the day and year in a large sample of healthy controls. We replicated previous findings of seasonal changes in the availability of the

serotonin 1A receptor, failed to replicate findings of seasonal changes in the availability of the serotonin transporter, and additionally showed diurnal changes in both targets.

In **Study V**, the importance of reliability is discussed with reference to study design, and a new method is presented for making approximations of the reliability for new samples. This approach allows researchers to more effectively gauge the feasibility of new between-individual studies before collection of any data, and to focus their efforts on research questions which can be expected to yield more interpretable outcomes.

In **Study VI**, we perform a direction replication of a previous finding of a strong association between the Self-Transcendence scale of the Temperament and Character Inventory using a much larger sample to assess the veracity of the original findings. We showed moderate to strong evidence for no effect relative to the the previous results, suggesting that the original results were more likely to be either a false positive or greatly overestimated.

In **Study VII**, we carried out an individual-participant data meta-analysis of TSPO binding measured using second-generation tracers in healthy controls compared with schizophrenia and psychotic disorder patients. Contrary to the original hypothesis of increases in TSPO binding, we showed strong evidence for decreases in TSPO in patients compared to controls in both cortical and subcortical regions.

In **Studies VIII and IX**, we hypothesised that D1 receptor binding would be higher with increasing proneness to develop psychosis and in the early stages of the disease prior to medication exposure, respectively. In **Study VIII**, we showed convergent evidence of no association between D1 receptor availability and delusional ideation in healthy controls. In **Study IX**, contrary to our hypotheses, we found moderate evidence in favour of lower levels of D1 receptor availability in the dorsolateral prefrontal cortex of first-episode drug-naive psychosis patients compared to healthy controls.

While reliability and replicablity of previous findings were directly assessed, the theme of reproducibility concerned our sharing the analysis code, and the data where possible, such that all analysis steps including those which could not be adequately described in the papers were recorded to ensure transparency, and demonstrate the correctness of our conclusions.

All of the three themes of the thesis concern efforts to improve the quality, robustness, and utility of scientific research. For a field such as PET imaging, which is not only resource-intensive, but also requires exposing participants to harmful radiation, it is especially important both from a scientific as well as an ethical perspective that data are processed and analysed in a manner which is transparent, generalisable and optimal such that they are made use of to their full potential.

List of publications

I **Matheson GJ**, Stenkrona P, Cselényi Z, Plavén-Sigray P, Halldin C, Farde L & Cervenka S (2017). Reliability of volumetric and surface-based normalisation and smoothing techniques for PET analysis of the cortex: A test-retest analysis using [11C]SCH-23390. *NeuroImage*, *155*, 344-353.

II Stenkrona P, **Matheson GJ**, Cervenka S, Plavén-Sigray P, Halldin C & Farde L (2018). [11C] SCH23390 binding to the D1-dopamine receptor in the human brain-a comparison of manual and automated methods for image analysis. *EJNMMI research*, *8*, 74.

III **Matheson GJ**, Plavén-Sigray P, Forsberg A, Varrone A, Farde L & Cervenka S (2017). Assessment of simplified ratio-based approaches for quantification of PET [11C] PBR28 data. *EJNMMI research*, *7*, 58.

IV **Matheson GJ**, Schain M, Almeida R, Lundberg J, Cselényi Z, Borg J, Varrone A, Farde L & Cervenka S (2015). Diurnal and seasonal variation of the brain serotonin system in healthy male subjects. *NeuroImage*, *112*, 225-231.

V **Matheson GJ** (2018). We need to talk about reliability: Making better use of test retest studies for study design and interpretation. *bioRxiv*, *274894*, *Under review*.

VI Griffioen G, **Matheson GJ**, Cervenka S, Farde L & Borg J (2017). Serotonin 5-HT1A receptor binding and self-transcendence in healthy control subjects-a replication study using Bayesian hypothesis testing. *PeerJ*, *6*, e5790.

VII Plavén-Sigray P, **Matheson GJ**, Collste K, Ashok AH, Coughlin JM, Howes OD, Mizrahi R, Pomper MG, Rusjan P, Veronese M, Wang Y, & Cervenka S (2018). Positron Emission Tomography Studies of the Glial Cell Marker Translocator Protein in Patients With Psychosis: A Meta-analysis Using Individual Participant Data. *Biological psychiatry*, *84*, 433-442.

VIII **Matheson GJ**, Plavén-Sigray P, Louzolo A, Borg J, Farde L, Petrovic P & Cervenka S (2018). Dopamine D1 receptor availability is not associated with delusional ideation measures of psychosis proneness. *bioRxiv*, *321646*, *Submitted*.

IX Stenkrona P, **Matheson GJ**, Halldin C, Cervenka S & Farde L. Reduced frontal dopamine D1-receptor binding in first episode drug naive patients with schizophreniform psychosis - a PET-study using [11C]SCH23390. *Manuscript*, *Submitted*.

List of additional publications

Borg J, Cervenka S, Kuja-Halkola R, **Matheson GJ**, Jönsson EG, Lichtenstein P, Henningsson S, Ichimiya T, Larsson H, Stenkrona P, Halldin C, Farde L (2016). Contribution of non-genetic factors to dopamine and serotonin receptor availability in the adult human brain. *Molecular psychiatry*, *21*, 1077-1084.

Plavén-Sigray P, Hedman E, Victorsson P, **Matheson GJ**, Forsberg A, Djurfeldt DR, Rück C, Halldin C, Lindefors N & Cervenka S (2017). Extrastriatal dopamine D2-receptor availability in social anxiety disorder. *European Neuropsychopharmacology*, *27*, 462-469.

Plavén-Sigray P[†], **Matheson GJ**[†], Schiffler BC[†] & Thompson WH (2017). The readability of scientific texts is decreasing over time. *eLife*, *6*, e27725.

Plavén-Sigray P, **Matheson GJ**, Gustavsson P, Stenkrona P, Halldin C, Farde L & Cervenka S (2018). Is dopamine D1 receptor availability related to social behavior? A positron emission tomography replication study. *PloS one*, *13*, e0193770.

Plavén-Sigray P, **Matheson GJ**, Cselényi Z, Jucaite A, Farde L & Cervenka S (2018). Test-retest reliability and convergent validity of (R)-[11C] PK11195 outcome measures without arterial input function. *bioRxiv*, *298992*, *Under review.*

[†] *Authors contributed equally to this work*

List of open source software contributions

Matheson GJ (2018). kinfitr: Kinetic Modelling of PET Time Activity Curves. R package version 0.3.1 *https://github.com/mathesong/kipettools*.

Matheson GJ (2018). relfeas: Reliability for Feasibility Analysis. R package version 0.0.1 *https://github.com/mathesong/relfeas*.

Padfield D & **Matheson GJ** (2018). nls.multstart: Robust Non-Linear Regression using AIC Scores. R package version 1.0.0. *https://github.com/padpadpadpad/nls.multstart*.

Talagala PD, Mancarci O, Padfield D & **Matheson GJ** (2018). staplr: A Toolkit for PDF Files. R package version 2.2.1 *https://github.com/pridiltal/staplr*.

Contents

Abbreviations

Abbreviation	Term
3D-OP-OSEM	3D-ordinary Poisson ordered subset expectation maximization
APD	Absolute Percentage Difference
AD	Alzheimer's Disease
AUC	Area under the curve
AIF	Arterial input function
ABSS	Automated blood sampling system
BF	Bayes factor
BP_X	Binding Potential relative to X compartment
BIDS	Brain imaging data structure
COV	Coefficient of variation
B_{avail}	Density of receptors available to bind in vivo
K_D	Dissociation constant
DVR	Distribution volume ratio
EEG	Electroencephalography
f_X	Free fraction in X compartment
FWHM	Full width at half maximum
fMRI	Functional Magnetic Resonance Imaging
HAB	High-affinity binder
HARK	Hypothesising after the results are known
IRF	Impulse response function
IPD	Individual participant data
ICC	Intraclass correlation coefficient
KI	Karolinska Institutet
LAB	Low-affinity binder
MRI	Magnetic Resonance Imaging
MRS	Magnetic Resonance Spectroscopy

MEG	Magnetoencephalography
MCMC	Markov chain monte carlo
MAB	Mixed-affinity binder
MNI	Montreal Neurological Institute
NHST	Null hypothesis significance testing
PET	Positron Emission Tomography
PCA	Principal component analysis
QRP	Questionable research practice
ROI	Region of Interest
5-HT	Serotonin
SRTM	Simplified reference tissue model
SPECT	Single Photon Emission Computed Tomography
SDD	Smallest detectable difference
SESOI	Smallest effect size of interest
SEM	Standard error of measurement
SUV	Standardised uptake value
SUVR	SUV ratio
TCI	Temperament and Character Inventory
TAC	Time Activity Curve
TCM	Tissue compartment model
TSPO	Translocator protein
WAPI	Wavelet-transform aided parametric imaging
WSCV	Within-subject coefficient of variation
V_X	Volume of Distribution of X compartment

Chapter 1

Introduction

The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.

– John W. Tukey

1.1 Positron Emission Tomography

Positron Emission Tomography (PET) is an *in vivo* biomedical imaging method which provides detailed information about biochemistry as well as functional and metabolic activity. The spatial resolution of PET is in the range of millimetres, while Magnetic Resonance Imaging (MRI) can achieve sub-millimetre resolution. PET imaging has a temporal resolution of several minutes to hours, while functional MRI (fMRI) can measure events of only a few seconds, and EEG or MEG recordings can measure electrochemical changes in the brain at a millisecond resolution[1]. However, while PET cannot compare to these other methods in the spatiotemporal domain, its strengths are in its biochemical sensitivity and specifity, which are unparalleled compared to other *in vivo* imaging methods. While magnetic resonance spectroscopy (MRS) using an MRI system can detect and quantify concentrations in the millimolar (i.e. 10^{-3}M) range, PET has the sensitivity to detect concentrations in the picomolar (i.e. 10^{-12}M range)[2]. The specificity of PET is related to its use of radiolabelled ligands which bind to target molecules of interest. These radioligands are designed such that they exhibit a high binding affinity for the target molecule which is ideally several orders of magnitude greater than for other related molecules. This means that the vast majority of the final signal should be

attributable to the availability of the specific target of interest. For this reason, PET imaging yields unique information from that provided by other *in vivo* imaging methodologies, and can thereby answer different questions.

1.1.1 Applications

PET imaging is used in clinical diagnosis, pharmaceutical research as well as research more broadly due to its properties of high biochemical sensitivity and specificity, as well as providing this information with detailed spatial resolution.

Common clinical applications of PET are in oncology, cardiology, and neurology. An example is the use of PET using radioligands targeting the β-amyloid deposition for the diagnosis of Alzheimer's Disease (AD). In this way, PET imaging can differentiate AD from other denemtia syndromes whose symptoms might be similar, but whose underlying pathology is different. A high degree of certainty can be gained from visual inspection of a measurement following even very simplistic quantification and short measurement duration.[3], as AD is associated with greatly increased signal intensity in several regions of the brain grey matter.

In pharmaceutical research, PET is often used in microdosing[4] and occupancy[5] studies. Microdosing studies involve radiolabelling the drug molecule itself and assessing its uptake into the body, allowing researchers to determine for example whether the drug can penetrate the blood brain barrier and enter the brain. Occupancy studies involve PET measurements with established radioligands, before and after the administration of different doses of the drug of interest. With these studies, researchers can determine whether the drug can bind (and thereby block) the target of interest, and the extent to which this occurs. These estimates can also be used to determine the degree of target occupancy associated with therepeutic effects[6], and side effects[7], and thereby select appropriate doses of the drug.

In research more broadly, and in the study of the brain specifically, PET is used to answer questions relating to brain physiology and to both psychiatric and neurological pathology. In the study of brain physiology, PET studies can be broadly categorised into those examining 1) the association of PET-derived outcome measures with theoretically stable attributes (such as personality traits[8] or cognitive performance[9]), 2) occupancy by endogenous neurotransmitter release or depletion[10–12] during or following acute behavioural[13] or pharmacological[14,15] interventions, or 3) changes in target concentration before and after prolonged behavioural[16] or pharmacological[17] interventions or simply in healthy ageing[18–20]. In the study of pathology, PET studies are generally focused either on patient-control comparisons[21–24], or on examining within-individual effects following treatment and their association to reductions in

symptoms[25,26].

There are a wide variety of experimental contexts for which PET imaging can be employed, and a diverse range of biochemical targets which can be studied owing to the extensive range of PET radiotracers available. As such, PET allows for the quantification of numerous physiological parameters, including neurotransmitter receptors and transporters, abnormal deposits of proteins and peptides, blood flow and volume, protein synthesis, metabolism of various substances, endogenous concentrations of neurotransmitters and their release, as well as the dynamics, kinetics and distribution of various drugs. From the perspective of measurement and study design, PET imaging is commonly used to make comparisons between individuals, within individuals, or even to compare within-individual changes between individuals.

1.1.2 Measurement

1.1.2.1 Radiochemistry

Radiotracers for PET imaging are synthesised by incorporating a radioactive isotope, commonly ^{11}C, ^{15}O or ^{18}F, into chemical molecules which bind to, or are taken up by, the target of interest. The isotopes are first created using a cyclotron. The labelling of the molecules is subsequently performed using precursors. This process is usually mostly automated to ensure that synthesis is rapid in order for radioisotopes to have decayed minimally before injection.

Following quality control, the radioligand is cleared for injection. There are several important properties of the final injected dose which are recorded. The *injected radioactivity* is proportional to the number of injected atoms of the radioactive isotope, and hence to the number of tracer molecules containing radioactive isotopes (i.e. "hot" tracer). It is typically measured in units of megabecquerel (MBq). The *injected mass* is, for a given tracer compound, proportional to the number of injected molecules. It is measured in units of micrograms (μg). The *specific activity* is the ratio of the injected radioactivity to the injected mass, which is a measure of the (small) proportion of injected tracer molecules which contain radioactive isotopes, compared to those which do not contain radioactive isotopes, i.e. "cold" tracer. This is typically expressed in units of megabecquerel per microgram (MBq/μg). This is similar to the *molar activity*, which is the fraction of injected radioactivity to the number of moles of the compound, typically measured in units of megabecquerel per micromolar (MBq/μmol). As such, synthesis of radiotracers with a high specific/molar activity allows for the injected compound to consist of a sufficient level of radioactivity, thereby ensuring sufficient count statistics during measurement in the PET system, without requiring injection of high mass, i.e. a large number of molecules of the

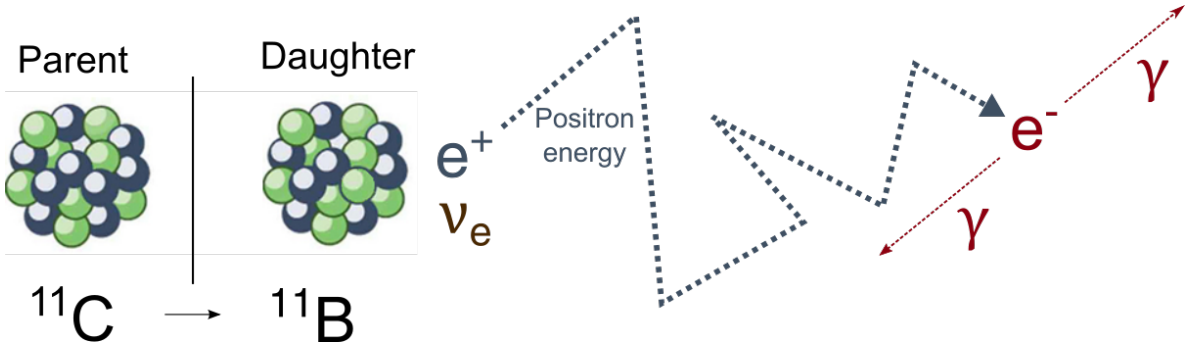

Figure 1.1: When PET radioisotopes decay, in this case carbon-11 to boron-11, a positron (e^+) and an electron neutrino (ν_e) are released. The positron collides with a nearby electron (e^-). This gives rise to the emission of two γ photons in approximately opposite directions. The figure is modified from Figure 31.4.5 from OpenStax College (2016) (27), licensed under CC BY.

cold compound which, in sufficiently large quantities, cause a small degree of blocking of the target.

1.1.2.2 Acquisition

Radioligands are usually administered through intravenous injection. PET radioisotopes undergo β^+ decay, meaning that one of the protons in the nucleus of the parent nuclide becomes a neutron, and a positron and a neutrino are emitted. The positron is emitted with a specific amount of kinetic energy (positron energy) dependent on the specific radionuclide. When this positron comes into contact with a nearby electron, usually within a millimetre of its release, the positron and electron annihilate together, producing a pair of γ photons, which travel in a random direction at approximately 180° to one another (see Figure 1.1).

These two photons travel through tissue, skull and air and eventually come into contact with detectors mounted in a ring around the subject (see Figure 1.2 A). These detectors consist of scintillating crystals, which emit visible light when γ particles pass through them. The crystals are coupled to a photomultiplier tube, which generates an electric response and registers a detection whenever the crystals scintillate. When two photons are detected simultaneously, this is recorded as a coincidence, and it is assumed that the positron-electron annihilation occurred at some position along the line of response between the two detectors (Figure 1.2 A in orange).

Some of these lines of response are detected are erroneous. Some of the γ photons are scattered by collisions with other molecules: while many of these events are not recorded due to the

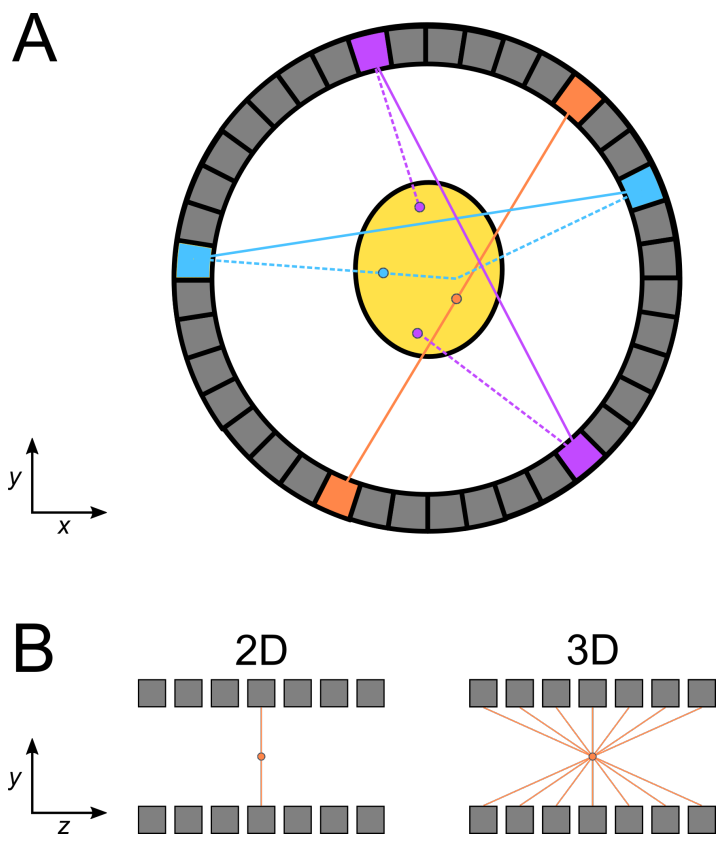

Figure 1.2: A. The simultaneous detection of two γ photons in two detectors is recorded as a line of response between the two detectors along which it is assumed that the annihilation took place. Small circles represent the annihilations, dashed lines represent the path of the photons, and the coloured detectors represent the detectors at which a scintillation took place. The orange line of response represents a correct recording, where the line of response and the path of the photons are overlapping. The blue line of response represents a scattered event. The purple line of response represents a random coincidence. B. The difference between 2D and 3D PET measurement and the permitted lines of response for a given location of the annihilation.

scattered photon never being detected (i.e. there is no coincident detection), if both photons are detected, then the line of response will be inferred incorrectly between the two detectors, called a *scattered event* (Figure 1.2 A in blue). Sometimes, by chance, two γ photons which were emitted from different annihilations will be detected simultaneously, resulting in a coincidence detection. This leads to an incorrectly inferred line of response, called a *random coincidence* (Figure 1.2 A in purple).

PET measurements can be recorded in two- or three-dimensions (see Figure 1.2 B). In 2D measurements, only coincident detections in the same or closely neighbouring rings of detectors (in the axial direction, i.e. head-to-toe) are recorded. PET measurements acquired in 3D have

a much larger number of counts, which corresponds with a higher sensitivity, however they are associated with a much higher number of both scattered and random events, and reconstruction is more computationally intensive[28,29]. While PET systems were initially operated exclusively in 2D, PET systems used in research today are mostly, if not exclusively, operated in 3D. This is due to the relative increase in computational power, the development and application of more sophisticated methods of reconstruction, as well as the use of sheilding to reduce random counts arising from activity outside the field of view of the camera[30].

1.1.2.3 Reconstruction

Following detection of a large number of lines of response, PET time frames can be reconstructed using algorithms which assign coincident detections to particular voxels of the brain. This is performed in such a way as to determine the most likely origin of each annihilation within the duration of the given time frame. Filtered backprojection is used in studies taking place with the HR PET system, while *3D-ordinary Poisson ordered subset expectation maximisation (3D-OP-OSEM)* is used in studies making use of the HRRT PET system[31,32]. While iterative methods such as 3D-OP-OSEM may exhibit bias for frames with low numbers of counts[33,34], and are more computationally intensive[35], they provide images with superior resolution, whose outcome measures are highly correlated with those of filtered backprojection[35–37].

The resolution of a PET image is its ability to spatially discriminate between different sources of radioactivity. As such, the resolution of PET is inherently limited by several factors. First, the number of detectors is important: with more detectors, lines of response are defined more accurately. Secondly, the distance between the release of the photon and the annihilation with an electron result in a degree of blurring of the image. Thirdly, the angle at which the photons travel away from one another is not quite 180° due to conservation of momentum, which also leads to blurring. Fourth, some lines of response are erroneous due to scattered events and random coincidences, leading to noise in the reconstructed image. Following reconstruction, and following adjustment for decay of radioactivity, a PET image is produced consisting of three dimensional frames, for which the value assigned to each voxel is an estimated concentration of radioactivity within the spatial confines of that voxel.

1.1.3 Quantification

1.1.3.1 Regions of the brain

Following reconstruction, a PET image is obtained, which is a 3D image consisting of estimated radioactivity concentrations in each 3D pixel (voxels) in the units of radioactivity per unit volume (typically kBq/ml). Static PET images consist of one 3D image acquired over the whole period of measurement, while dynamic PET images consist of a series of 3D PET images acquired during different time frames. The series of radioactivity concentrations within each voxel, or the mean radioactivity concentration within a set of voxels (referred to as a region of interest, ROI) over time is referred to as a *time activity curve (TAC)*. For the purpose of defining which voxels are of anatomical interest, and which should be combined to form specific anatomical ROIs, PET images are coregistered to anatomical T1-weighted (anatomical) MR images acquired for the same individuals. Anatomically relevant regions of the brain can be defined on the MR image, and resliced to the space of the PET image to define these regions.

While comparison of outcome measures in ROIs is more commonly used in PET research, when examining effects for which there is no strong a priori regional hypothesis, or when fine-grained regional heterogeneity of the effects is expected, exploratory spatial methods represent a useful approach. These methods involve calculation of parameters of interest within each voxel of the brain independently, in contrast to ROI-based approaches. They are therefore referred to as parametric images. This approach is therefore more computationally intensive than ROI-based approaches as there are more voxels than ROIs, but also substantially more prone to error in calculation of the parameters of interest due to the noisy reconstruction of each individual voxel. A further complication is that, in order to make comparisons between individuals, the spatial positions of these voxels must be comparable. This involves transforming the parametric image of the brain into a common set of spatial coordinates. Statistical comparisons between individuals, groups or conditions are thereby made at each voxel of this shared spatial coordinate system independently, although this necessitates an increased type I error rate. Spatial smoothing is applied in order to increase the signal-to-noise ratio of the resulting spatial parameter estimates, either prior to, or following kinetic modelling. Smoothing also partially compensates for the problem of multiple comparisons, allowing the use of Random Field Theory to perform familywise error rate correction at the level of clusters rather than individual voxels[38].

For parametric analysis of cortical regions, the surface-area-to-volume ratio of the grey matter is very high. This means that three-dimensional Gaussian smoothing can lead to problems as a large proportion of the signal within cortical voxels following smoothing can have originated

outside of the cortical grey matter. Additionally, signal from neighbouring sulci can be smoothed into one another, despite their being far from one another within the cortical grey matter itself. More recently, surface-based registration and smoothing methods have been developed, by which values are first sampled from the centre of the cortical ribbon to a cortical surface, following which registration and smoothing are performed using only these values in two dimensions[39,40]. In this way, all signal originates in cortical grey matter, and registration and smoothing are based on cortical folding patterns, and are thus able to be more anatomically precise than volumetric methods allow. It was shown by Greve et al. (2014)[40] that surface-based registration and smoothing led to greatly reduced intersubject variance and bias, and could allow for greater statistical sensitivity in applied clinical studies using parametric analysis.

1.1.3.2 Kinetic Modelling

From a TACs describing the time course of radioactivity concentration in the ROI, researchers usually seek to derive scalar estimates describing the concentration of the target of interest (or some proxy thereof) based on all of these values. Quantification of these outcome measures involves comparing the radioactivity concentration in the target with a reference radioactivity concentration such as the arterial plasma (such as the *total volume of distribution,* V_T) or a reference region (such as *binding potential relative to the non-displaceable compartment,* BP_{ND}) of the brain using a kinetic model. Alternatively, the radioactivity in the target of interest can be compared with the injected radioactivity per unit body mass, or the background signal during the course of measurement for semi-quantitative estimates such as the *standardised uptake value (SUV).*

The objective of kinetic modelling is to derive an analytical expression which describes the measured TAC, and to use this mathematical description to estimate outcome measures which describe the concentration of the target in the given anatomical location, or to describe the behaviour of the radiotracer in the tissue. These outcome measures can then be statistically compared between individuals, groups or conditions for a given region of the brain.

Compartmental models are the classical form of kinetic modelling. This means that the concentration of the radiotracer within the target tissue and the reference input function (whether it may be brain tissue or arterial plasma) are described according to compartments between which the radiotracer can be transferred. The transfer between compartments is described using rate constants. Models making use of the metabolite-corrected arterial plasma as input function (shortened to the *arterial input function, AIF*) or are based on the three-tissue

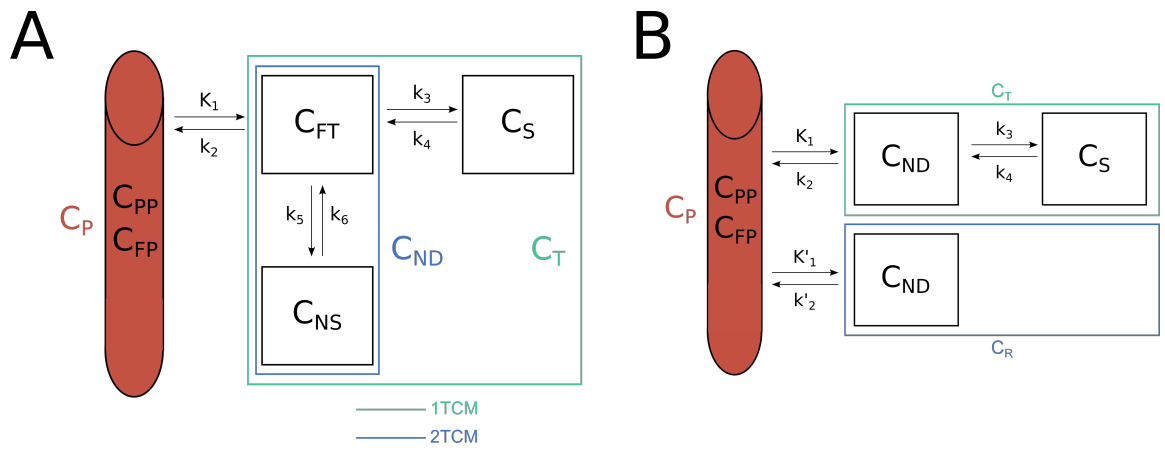

Figure 1.3: For both panels, C represents the radioactivity concentrations within each compartment. The red cylinder on the left of each panel represents the artery, containing plasma (P). Within the plasma, the radiotracer is either free (FP), or bound to plasma proteins (PP). The black boxes represent the compartments. TCM refers to Tissue Compartment Model. A. The three tissue compartment model is the basis for the two- and one-tissue compartment models: transfer between certain compartments are assumed to be sufficiently rapid that they can be considered as single compartments for the two- and one-tissue compartment models (coloured boxes). The compartments include FT free tracer, NS non-specifically bound, S specifically bound, T total, and ND non-displaceable. B. Reference region models consider the total concentration of radiotracer in the target T and in the reference region R, and assume that the non-displaceable concentration is comparable in both regions, and that the specific binding in the reference region is equal to 0.

compartmental model and simplifications thereof (Figure 1.3A), while models making use of reference regions compare the concentration of radiotracer in the target and reference regions under the assumption that the specific binding in the reference region is equal to zero, or is at least negligible (Figure 1.3B).

Full quantification methods yield two classes of outcomes for reversible tracers: *volumes of distribution (V)* and *binding potential (BP)*. Volumes of distribution define the concentration of the target protein in different compartments of the target tissue (V_T total, V_S specific, V_{ND} non-displaceable, V_{NS} non-specific, V_{FT} free tracer) relative to a common reference: the concentration of radiotracer in the arterial plasma (including that which is bound to plasma proteins)[41].

$$V_T = V_{FT} + V_{NS} + V_S = V_{ND} + V_S$$

Volumes of distribution can alternatively be described in terms of rate constants as follows:

$$V_T = \frac{K_1}{k2}(1 + \frac{k_3}{k_4})$$

$$V_{ND} = \frac{K_1}{k_2}$$

$$V_S = \frac{K_1 k_2}{k_3 k_4}$$

Binding potential (BP) values have in common that they are defined as the concentration of target protein in the specifically bound compartment, relative to different reference concentrations (BP_{ND} non-displaceable compartment, BP_P plasma, or BP_F the free concentration in plasma). BP can also be conceptualised in terms of *in vitro* measures of B_{avail}, the concentration of the available unbound target, K_D, the dissociation constant, and the free fraction in either the non-displaceable compartment (f_{ND}), or the plasma (f_P)[41].

$$BP_{ND} = \frac{V_T - V_{ND}}{V_{ND}} = \frac{V_S}{V_{ND}} = f_{ND}\frac{B_{avail}}{K_D} = \frac{k_3}{k_4}$$

$$BP_P = V_T - V_{ND} = f_P\frac{B_{avail}}{K_D} = \frac{K_1 k_2}{k_3 k_4} = V_S$$

$$BP_F = \frac{V_T - V_{ND}}{f_P} = \frac{BP_P}{f_P} = \frac{B_{avail}}{K_D} = \frac{K_1 k_2}{f_P k_3 k_4}$$

These are all examlples of full quantification methods: these differ from semi-quantitative estimates in that they compare the concentration of radioactivity in tissue not just during the course of the measurement (i.e. *their area under the curve, AUC*), but their integral to infinity, i.e.

$$V_T = \frac{\int_0^\infty C_T(t)dt}{\int_0^\infty C_P(t)dt}$$

An example of a semi-quantitative methods is the standardised uptake value (SUV), which makes use of the AUC during a specified time window, and compares this to a reference quantity, in this case the ratio of the injected dose per unit body mass. Another semi-quantitative method is the SUV ratio, which is a measure of the ratio of SUV values from two different

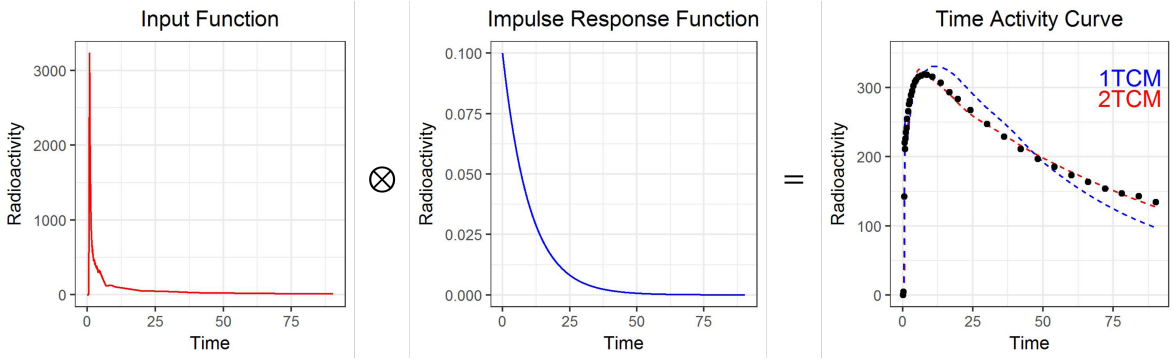

Figure 1.4: The TAC can be described as the convolution of the input function with the IRF. 1TCM and 2TCM refers to the fits of the one- and two-tissue compartment models.

regions.

The instantaneous rate of change of C_T can be described by flow of radiotracer to and from the plasma, i.e. $\frac{dC_T(t)}{dt} = K_1 C_P(t) - k_2 C_T(t)$. Solving the differential equation reveals that the concentration of C_T at each point in time, i.e. the TAC, can be described by a convolution of the AIF with the *impulse response function (IRF)* (Figure 1.4). The IRF is the function which describes the concentration of the radiotracer over time if the delivery of the radiotracer were instantaneous (i.e. a Dirac delta function). The IRF therefore describes the efflux of the tracer from the tissue once it is already there. The TAC can be conceptualised as resulting from a continuous series of deliveries of the tracer over time as described by the input function, each of which flows out of the tissue according to the IRF. In the example in Figure 1.4, after 25 minutes, only a small proportion of the tracer which arrived during the first minute after injection remains in the tissue, while the small amount of tracer still present in the plasma at this point is still being delivered to the tissue.

Performing blood sampling to derive an AIF can cause discomfort for research participants, is resource-intensive due to both instrumentation and labour, and the measurements themselves are prone to measurement error. Alternatives to blood sampling are therefore desirable. Reference region models such as the *simplified reference tissue model (SRTM)*[42] do not require blood sampling, and provide estimates of BP_{ND} (for reversible tracers), provided that there is a region of the brain with zero or negligible specific binding. BP_{ND} can be calculated by assuming equality of V_{ND} between the target and the reference region, and by making use of the TAC of the reference tissue as input instead of the AIF.

Fitting kinetic models to data is typically performed by using nonlinear least squares estimation, by which rate constants describing the IRF are updated over successive iterations, and convolved with the input function until the created TAC matches the measured TAC (i.e. it minimises

the weighted sum of squared residuals). While these models are unbiased, they are not guaranteed to find the optimal solution, i.e. they may arrive at a local minimum, rather than the global minimum solution. They are also much more computationally intensive than linear models. While computational time is barely relevant for the fitting of individual TACs, this becomes a significant issue for the fitting of multiple TACs at the level of voxels. Furthermore, TACs at the voxel level are usually much noisier than for ROIs, and therefore convergence issues are more likely for nonlinear least squares estimation. For this purpose, linearisations of kinetic models have been developed, which enable the use of linear least-squares fitting procedures, which are much quicker, and guaranteed to arrive at the global minimum. These models typically make use of asymptotic properties of the tissue response and provide only the macroparameters (e.g. only V_T or BP_{ND} values for invasive or non-invasive models respectively, and not rate constants, i.e. microparameters). There exist numerous linearised models for the fitting of data using both plasma input functions (e.g. Logan graphical analysis[43]) as well as reference tissue models (e.g. the non-invasive Logan plot[44]). However, these models may often provide slightly biased outcomes[45], but may also reduce noise (variance) in the process.

Individual rate parameters, or microparameters, are estimated by compartmental models. These rate constants can be compared between individuals, or used to obtain measures of binding potential directly (e.g. $BP_{ND} = \frac{k_3}{k_4}$), however this approach is prone to noise and usually produces unreliable estimates compared to calculation of the more conservative V_T[46]. This is because estimation of the rate constants may often suffer from a lack of identifiability. This means that they can not be estimated uniquely from the data, and that multiple different combinations of values of the rate constants, which may differ considerably, can give rise to a very similar predictions of the measured curve (Figure 1.5). The identifiability of these outcomes is determined in part by the degree of statistical noise in the measured signal[46].

There is no one optimal model or strategy for quantification of PET TAC data. Selection of the appropriate model should be performed by taking into consideration not only the radioligand and its kinetic behaviour, but also the results of previous test-retest evaluations, the research question, as well as practical considerations for performing the study.

1.2 Statistical Inference

Most biomedical research is performed using the null hypothesis significance testing (NHST) paradigm[47]. These approaches are referred to as frequentist as they concern themselves with p values, which represent frequencies. In frequentist statistics, any given experiment can be considered as one of an infinite series of possible repetitions of the same experiment, each of

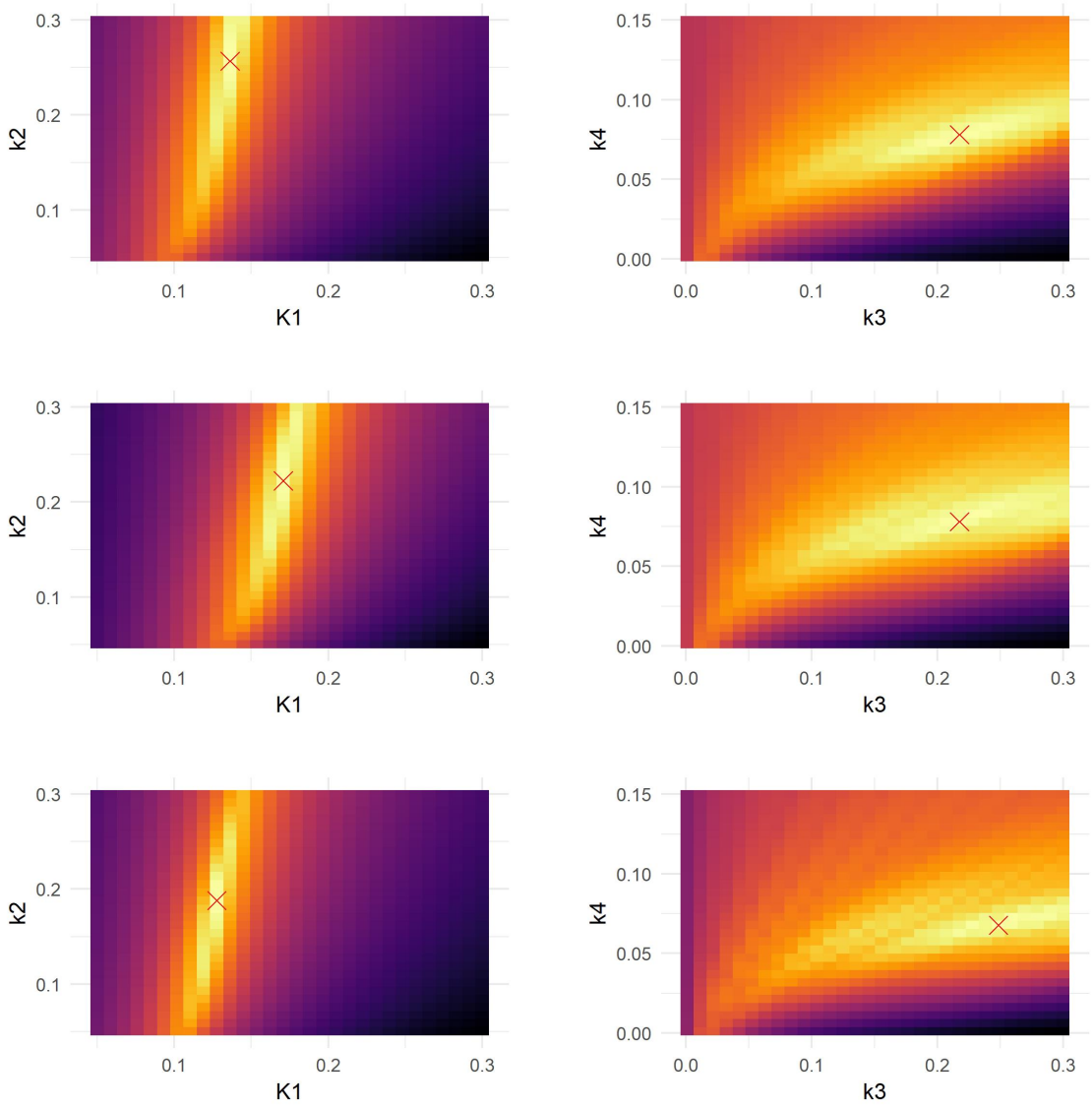

Figure 1.5: Identifiability of parameters relates to the degree to which the measured data, i.e. the time activity curve, can be described by a unique combination of the fitted parameters. In this figure, a single time activity curve from the same region from three individuals (rows) was fitted using a grid of 30 parameter values for each of the four rate constants, and the weighted sum of square residuals was calculated for each fit (i.e. the cost function). Each point represents the value of the best fit based on the other two rate constants for the specific combination of the rate constants represented by the x and y axes. In each case, the minimum is represented by a red cross. It is clear that several combinations of rate constants are associated with similar values of the cost function, implying that the goodness of their specific fit to the measured data is similar, and that identifiability of the individual rate constants is poor: some more poor than others. Notably, the *relationships between the parameters* appears to be somewhat more identifiable than for the individual rate constants themselves, and more so for V_{ND} (left) than for BP_{ND} (right).

which produces a statistically independent result. The outcomes of these infinite repetitions are calculated by what can be conceptualised as performing a simulation (or calculating the results of such a simulation analytically given certain assumptions). This simulation concerns what the distribution of outcomes would be if there were no effect at all in the repetitions of the same experiment, based on the observed data which constitutes a single repetition. The p value represents the proportion of the simulations which are more extreme than the obtained value from the study. The significance level is defined as a reference point beyond which only a prespecified proportion (e.g. 5%) of simulated repetitions of the experiment obtain effects which are more extreme, despite there being no true effect. If the p value observed is more extreme than the significance level, then it is deemed sufficiently incompatible with the null hypothesis. This approach is referred to as "proof by contradiction"[48,49].

The Fisherian approach uses the p value to assess how surprising the outcome is under the assumption of the null hypothesis being true. Smaller p values are therefore interpreted as being more surprising. The crux of the Neyman-Pearson approach is to define an alternative hypothesis and to frame the interpretation of p values in terms of error rates[47–49]. In the latter approach, the outcome of the experiment is dichotomised into rejecting or failing to reject the null hypothesis based on whether the p value is above or below a prespecified threshold. This means that the specific p value obtained is irrelevant, but that only whether it is above or below the threshold is considered[50]. As described before, based on the long-run (imaginary, or simulated) frequencies, the null hypothesis will be rejected 5% of the time when it is in fact true. This is referred to as a Type I error, i.e. a false positive. These errors are controlled by defining a Type I error rate, called α. The other type of error is defined as Type II error, i.e. a false negative. The type II error rate (β) is therefore the rate of failing to reject the null hypothesis when the null hypothesis is in fact false. This can alternatively be conceptualised as power: the rate of successfully rejecting the null hypothesis when the null hypothesis is false, i.e. $1 - \beta$. The power of an experiment is a function of the sample size and the the true underlying size of the effect (Figure 1.6). Because the latter quantity cannot be known, power analysis, and thereby sample size determination, is usually performed based on an estimated effect size. A disadvantage of this approach is that optimistic effect size estimates lead to optimistic estimates of power. An alternative, better strategy is to perform a study which has sufficient power to detect the smallest effect size of interest, such that for all interesting underlying true effect sizes, the null hypothesis will be reliably rejected[51].

This approach is not intuitive, and misinterpretations of p values are common[52], even by statisticians[53]. This has even prompted the American Statistical Association to issue a statement urging scientists not to simply rely on whether such a summary statistic passes a given threshold, but to take results in greater context[54]. Common misconceptions about p

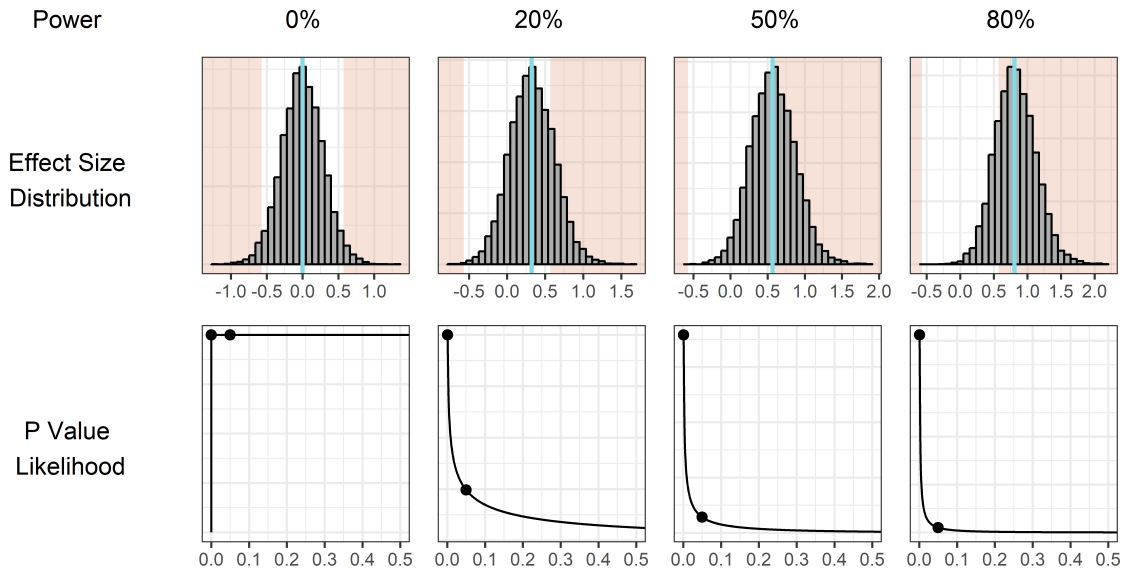

Figure 1.6: Power is a function of the true underlying effect size and the sample size. In the upper row, the observed effect sizes obtained by simulation are shown by the histogram, with the underlying, true, effect size depicted by the blue line. The width of the observed effect size distribution is a function of the sample size i.e. larger samples yield effect size estimates with lower variability. The transparent red region depicts those outcomes which are statistically significant. The power represents the proportion of the observed effect sizes which are significant, i.e. for 50% power, 50% of the observed effect sizes will be significant, but the other 50% are not significant due to type II error. For 0% power, i.e. the null hypothesis is true, 5% of the effect sizes are statistically significant due to type I error. In the lower row, the distribution of the likelihood of the resulting p values is shown, with dots depicting the relative likelihood of obtaining p values equal to 0.001 and 0.050. For 0% power, i.e. no effect, these values are equally likely and 5% of the effect sizes will be significant. For true effects, i.e. power>0%, p values of 0.001 are more likely to be observed than p values of 0.050 by the following margins: 4× for 20%, 13× for 50%, and 41× for 80% power.

values include the *inverse probability fallacy* (the belief that the p value indicates the probability that the null hypothesis is true), the *replication fallacy* (the belief that the p value reflects the replicability of a result), the *effect size fallacy* (the belief that the p value provides direct information about the effect size) and the *clinical or practical significance fallacy* (the belief that the p value relates to the importance of an effect size)[52,55]. It has also been suggested that a source of the misinterpretation of p values is that the inverse probability fallacy reflects the question that we, as scientists, are typically asking, i.e. what is the probability that the alternative hypothesis (i.e. often our research hypothesis) is true[56]? Following the growing concern around the replicability crisis in psychology and biomedical science[57–62], there have recently been suggestions to lower the default alpha threshold[63], not to rely on default alpha

thresholds[64], or even to abandon significance testing entirely[65].

Another issue is that of the null hypothesis: as described, a p value permits concluding either to reject the null hypothesis or to fail to reject the null hypothesis due to insufficient evidence. This leaves no room for being able to accept the null hypothesis, i.e. to say that there is no effect. However, when it is important for scientists to be able to conclude that a meaningful effect is absent, making this conclusion based on a nonsignificant p value is neither satisfactory nor statistically valid, although the practice is common[66].

There do, however, exist several approaches which can assess the null hypothesis in a statistically valid manner. I will describe two of these approaches which are utilised in this thesis. Within the frequentist tradition, equivalence testing involves defining a smallest effect size of interest (SESOI) as opposed to using a point estimate null hypothesis of zero[67,68]. The procedure can internally be as straightforward as performing two one-sided t tests against the SESOI bounds. This method thereby inverts the null hypothesis, and can reject a new null hypothesis that the effect size is larger than these bounds. Another alternative is the use of Bayesian statistics, which, due to defining prior probability distributions over each of the parameters of the model, is able to provide more nuanced conclusions than frequentist methods are usually capable of inferring.

In Bayesian statistics parameters are described by probability distributions over potential values. In this way, they differ from frequentist methods which concern themselves with the long-run frequencies of observing certain outcomes. Rather, with Bayesian statistics, inference can be performed with reference to the probabilities of parameters falling within certain ranges, or by comparing the relative likelihood of the data arising under two or more different hypotheses.

Broadly, Bayes' rule defines the optimal manner by which beliefs can be updated given new data[69,70]. In practice, when applying Bayesian inference, Bayes' rule allows for conversion between the probability of the data ($\mathcal{D}$), and the probability of the parameters (θ) under certain circumstances.

$$\underbrace{P(\theta|\mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{P(\theta)}^{\text{Prior}} \times \overbrace{P(\mathcal{D}|\theta)}^{\text{Likelihood}}}{\underbrace{P(\mathcal{D})}_{\substack{\text{Marginal} \\ \text{likelihood}}}} \tag{1.1}$$

This is useful, as it allows for describing the likelihood of a set of parameters conditional on (meaning given) the data (i.e. $P(\theta|\mathcal{D})$, the posterior), or for describing the likelihood of the

data conditional on a set of parameters (i.e. $P(\mathcal{D}|\theta)$, the likelihood). Bayesian modelling concerns itself with the posterior probability, while frequentist statistics concerns itself with the likelihood.

In order to be able to calculate the posterior from the likelihood, a prior distribution must be defined over the parameter values. This is occasionally viewed as a drawback, as it necessitates an increased degree of subjectivity. However, this can also be viewed as a strength, as we almost always have at least some knowledge about what values of the parameter are reasonable and unreasonable.

Inference in Bayesian statistics can take several different forms. *Parameter estimation* refers to examination of the posterior probability distribution for any the parameters, summarising it by its mean value and credible intervals. In this way, as described before, we can describe the interval in which there is a given percentage probability that the true parameter lies, given our prior beliefs and the new data by which to update those beliefs. Similarly, we can calculate the probability that a given parameter is above or below a certain value such as zero (called a posterior p value). *Bayesian hypothesis testing* is another alternative, by which multiple hypotheses ($\mathcal{H}$), described by models, are compared. The outcome of these tests is the Bayes Factor (BF) (equation 1.2).

$$\underbrace{\frac{P\left(\mathcal{H}_1|\mathcal{D}\right)}{P\left(\mathcal{H}_2|\mathcal{D}\right)}}_{\substack{\text{Posterior}\\ \text{model odds}}} = \underbrace{\left(\frac{P(\mathcal{D}|\mathcal{H}_1)}{P(\mathcal{D}|\mathcal{H}_2)}\right.}_{\text{Bayes factor}} \quad \underbrace{\left.\frac{P\left(\mathcal{H}_1\right)}{P\left(\mathcal{H}_2\right)}\right.}_{\substack{\text{Prior}\\ \text{model odds}}} \tag{1.2}$$

BFs can be interpreted in several different ways depending on the context, all of which are correct. BF_{12} represents the comparison of model 1 with model 2, such that $BF_{12}>1$ means that hypothesis 1 is preferred over hypothesis 2. BFs describe the predictive adequacy of competing models, or hypotheses, relative to one another. They can also be conceptualised as the relative probability of the observed data under one hypothesis compared to another[69,71], i.e. $BF_{12}=5$ means that the observed data are five times more likely to have occurred under hypothesis 1 compared to hypothesis 2. The posterior odds refers to the probability of one hypothesis relative to another, and is equal to the BF if both models were deemed to be equally plausible *a priori* (i.e. equal prior model odds) (equation 1.2). As such, the BF can also be conceptualised as the degree to which a skeptical observer can favour hypothesis 2 *a priori* before they should start favouring hypothesis 1, i.e. a $BF_{12}=5$ should influence an observer who previously believed that hypothesis 2 was up to 5 times more likely than hypothesis 1, to revise their beliefs and start to believe that hypothesis 1 is more likely than hypothesis 2. In this way, the BF functions as a belief update factor.

Within a Bayesian framework, the common misinterpretations made within frequentist statistics are often the correct interpretations of similar outcomes. For instance, confidence intervals within the frequentist paradigm do *not* represent the range in which there is a 95% probability that the true value lies[72] (although this misconception is widespread[73]), while this is the correct interpretation of credible intervals, their Bayesian analogue. This is an example of the inverse probability fallacy, which is not a fallacy within the Bayesian approach. Indeed Fisher described Bayesian statistics as "inverse probability"[74]. Similarly, in Bayesian hypothesis testing, one directly evaluates and compares the likelihood of different hypotheses against one another, while frequentist methods simply evaluate whether the null can be rejected. A further example is that of sequential testing: using frequentist statistics, researchers cannot decide to collect more data after checking whether the results were significant without performing correction for multiple comparisons. This is due to the use of error rates to constrain inference, since there are more opportunities to make errors. However, re-examination of the data, and collection until the evidence is sufficiently compelling, is of no consequence for Bayesian methods as they make use of probability updating[75].

1.3 Challenges for Biomedical Research

Before individual observations can be represented by scientific theories and thereby become knowledge, there are significant issues which must be overcome. These issues may either prevent scientists from gaining conclusions from observations, or worse, may mislead scientists into drawing incorrect conclusions from their observations. In other words, these issues concern the 'truth' of individual research claims themselves, and threaten the link between observations from research and science itself. A famous, and infamous, theoretical paper by Ioannidis (2005)[76] titled *Why most published research findings are false*, describes how current practices in the field of biomedical research lead to a large number of research findings which are untrue. Based on the considerations of the paper, several corollaries are presented, defining properties which would be expected to diminish the likelihood of truth of any given research finding. Of these, several can be said to be of particular relevance to the field of PET research, and especially its clinical application. These include studies being conducted with small sample sizes, examining small effect sizes, with flexibility of analysis outcomes, and with less a priori selection of tested relationships. It is therefore of critical importance for biomedical research, including clinical PET research, that the following issues be taken into consideration and addressed.

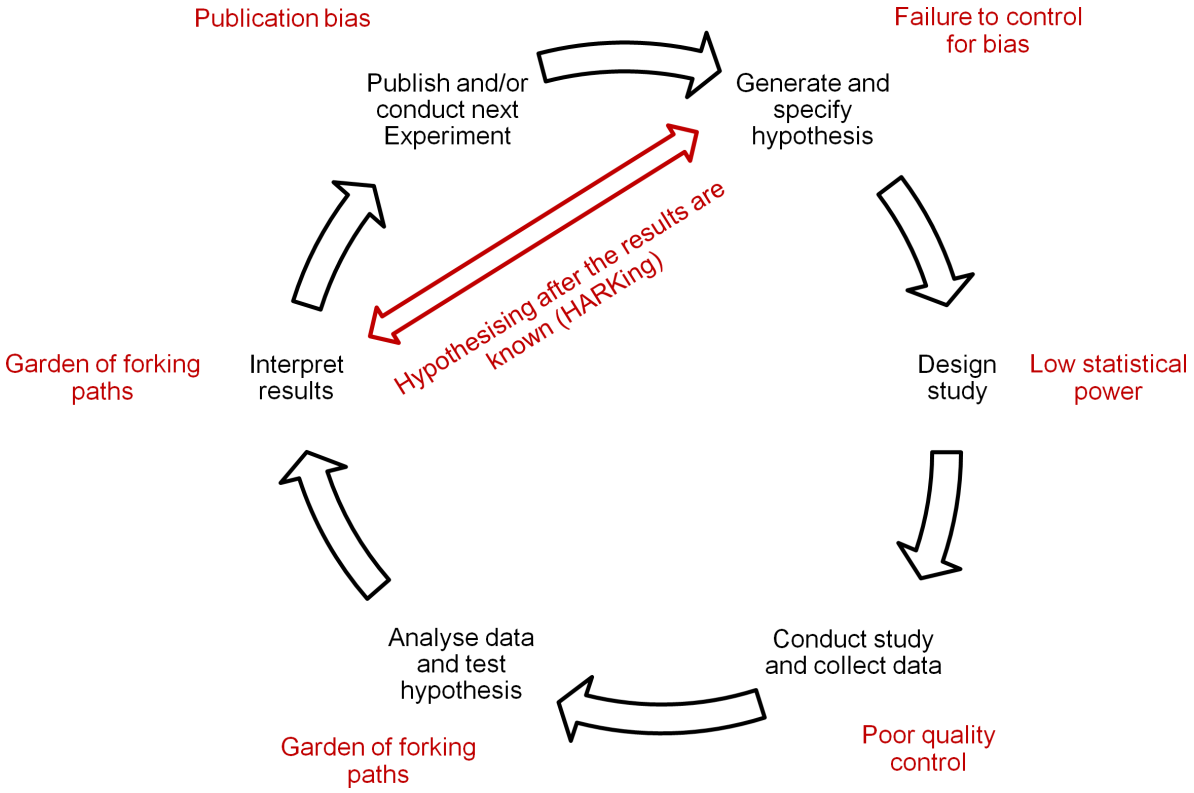

Figure 1.7: Threats to the veracity of scientific findings can occur at numerous different points across the hypothetico-deductive scientific method. In each case, the stage of research is presented in black, and the threat is presented in red. The figure is based on Figure 1 of Munafo et al. (2017) (77), licensed under CC BY.

1.3.1 Errors and Bias

There is a potential for errors and biases to occur throughout all stages of the scientific process (Figure 1.7) which should be considered. Below, I will expand upon most of these issues and how they relate to one another.

Low statistical power (Figure 1.7) implies that there is an insufficiently large set of data for providing a consistent answer to the specific research question. As described before (shown in Figure 1.6), the power is a function of the true, unknown, effect size and the sample size. Due to the high cost of PET examinations (individual measurements can cost more than USD 10 000), and the exposure of participants to harmful radioactivity, sample sizes are usually low. This would not be a problem if effect sizes were large (e.g. quantifying the dopamine transporter in Parkinson's Disease), however truly large effect sizes are unusual in psychiatric PET studies. As such, this issue is of particular relevance for this field, and particular efforts should be devoted to solving this issue (discussed more in the *Future Perspectives* chapter).

The issue of low statistical power is not unique to PET however, and rather appears to be a common feature of neuroscience research across numerous methodologies[78], and likely to most fields of biomedical research.

The act of analysing and interpreting data in several different ways, and settling on a method which produces the desired or expected outcome while neglecting to report the results of the other paths, has been likened to a *garden of forking paths*[79,80]. This has also been called P-hacking, but it has been argued that this latter term incorrectly implies an intention to cheat, while it is much more commonly performed unconsciously due to all of the above biases. It has been suggested to stem in part from the inverse probability fallacy: if one interprets a p value as the probability that the alternative hypothesis is true, then it would be less relevant that another method produced an insignificant result. It is only by correctly interpreting p values as error rates that it becomes clear that there have been more opportunities by which to make errors while exploring the different paths. Similarly, *publication bias* refers to publishing only 'successful' outcomes, and neglecting to publish those which did not show the expected or desired result. Publication bias, in contrast to the garden of forking paths, usually refers to entire studies. In both cases, however, neglecting to publish outcomes which were not statistically significant results in a substantial exaggeration of the average effect size based on the published literature (in meta-analyses for example[81]). In fact, with low power, *only* those studies (or forking paths of the analysis) whose observed effect sizes are exaggerated compared to the true underlying effect size will be statistically significant as can be seen from Figure 1.6.

Another related bias is that of *HARKing*, or hypothesising after the results are known[82]. Due primarily to hindsight bias, exploratory results are often seen as more predictable than they really were. This has also been described as just-so storytelling: finding a story to tell to rationalise the results[83]. Related again to poor statistical literacy among researchers, it is not commonly known that, in a strict Neyman-Pearson framework, p values apply exclusively to confirmatory research (i.e. hypotheses which were predicted in advance), and cannot be interpreted when they are used to explore the data from which effects are found (referred to as postdiction)[84]. To find a hypothesis among the final results, and to present a research finding as if this were the primary research question as predicted from the literature, leads to inflated effect sizes and increased rates of false positives[85].

The garden of forking paths, publication bias and HARKing are referred to as Questionable Research Practices (QRPs)[86]. These QRPs affect the planning of new studies, which, if they do not account for the sometimes dramatic positive bias in reported outcomes, will be overoptimistic about the expected underlying effect sizes. This may lead to selection of small sample sizes, and thereby low statistical power for the true underlying effect in future studies.

If all these practices were uncommon, then the degree of bias in the scientific literature would not be expected to be large. Studies of QRPs within the fields of psychology[86] and ecology[87] reveal, however, that they are common and widespread, suggesting that the bias within the scientific literature is likely to be severe.

While this leads us to conclude that a large proportion of the existing literature is quantitatively, and presumably even qualitatively, biased, there are no shortage of solutions for how this situation can be remedied in future, most of which relate to transparency, both at the stages of reporting and even as early as study planning; and openness of analysis methods, data, and materials[77]. This will be covered in more detail in later sections.

Another very recent encouraging result comes from a large-scale replication analysis of 21 social science papers published in Nature and Science between 2010 and 2015[88]. Sample sizes were defined such that they had 90% power to obtain an effect size of half the magnitude of the original findings, leading to sample sizes of on average five times the original sample sizes. While 62% of these effects replicated, what was most encouraging was a secondary analysis in which independent groups of peers were asked to predict the results of the replications through a prediction market before knowing the outcomes. The results of these predictions show perfect separation of predictions for studies which did, and did not, replicate in the correct directions, i.e. all studies which successfully replicated were predicted as being more likely to replicate than all studies which did not successfully replicate. This means that the replicability of research claims is unlikely to be simply random. Rather, research peers appear to be highly capable of detecting which results are "too good to be true"[88]. With increasing awareness of the above biases, and an increasing awareness that rigorous statistical training is important, the quality of research is likely to increase.

1.4 Reliability, Replicability and Reproducibility

The work within this thesis is centred around the concepts of reliability, replicability and reproducibility. In this section, these terms will be defined, and related to their metaphorical imagery on the front cover.

1.4.1 Reliability

Reliability, broadly, refers to the consistency of a measurement. In PET imaging, there are several different measures of consistency, which I will outline briefly before returning to

reliability specifically.

1.4.1.1 Coefficient of Variation (COV)

The coefficient of variation relates the dispersion, or spread of a set of observations, as a function of the mean value.

$$COV = \frac{\sigma}{\mu} \tag{1.3}$$

where σ represents the standard deviation, and μ represents the mean value of the sample. The COV is the inverse of Cohen's D: for this reason, a reduction of the COV results in greater power to detect an effect of a given proportional magnitude of mean difference. COV only applies for scales with an absolute 0.

1.4.1.2 Measurement Error

Each measurement is made with an associated error. This error can be described by its standard error (σ_e), which can be thought of as a standard deviation around the measured value referring to the spread other potential measurements of the same individual. It can be estimated from a set of measurements in which each individual is measured more than once using the following equation from Baumgartner et al. (2018)[89].

$$\sigma_e^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij} - \bar{y}_i)^2 \tag{1.4}$$

where n represents the number of participants, i represents the subject number, j represents the measurement number, k represents the number of measurements per subject, y represents the outcome and $\bar{y}_i$ represents the mean outcome for that subject.

1.4.1.3 Repeatability

The repeatability of an observation refers to the closeness of agreement between successive measurements[90]. In the field of PET imaging, this is often described using the absolute percentage difference (APD), or absolute variability, which relates the within-individual changes in the outcome value to its mean.

$$APD = \frac{|y_2 - y_1|}{\frac{1}{2}(|y_1 + y_2|)} \tag{1.5}$$

where y represents the outcome, and the subscripts refer to the measurement number. The average APD across the sample of test-retest measurements is usually presented to describe the repeatability for a particular outcome. Baumgartner et al. (2018)[89], however, points out that this measure is not sufficiently sensitive to outliers whose two measures might be highly inconsistent. Instead, the within-subject coefficient of variation (WSCV) is recommended[91], which relates the within-subject variation to the mean outcome across the group.

$$WSCV = \frac{\sigma_e}{\mu} \tag{1.6}$$

This measure can easily be converted to the smallest detectable difference (SDD), or repeatability coefficient, which describes the smallest within-individual change which could be considered sufficiently large that it is sufficiently unlikely to occurred by chance alone (according to a given confidence interval, e.g. using $z_{(1-\alpha/2)}$=1.96 below)[89,92].

$$SDD = \sqrt{2} \times z_{(1-\alpha/2)} \times \sigma_e \tag{1.7}$$

The WSCV and SDD are relevant for understanding the expected degree of variability within individuals which can be expected due to random fluctuations, and can therefore be used the gauge the feasibility of within-subject study designs.

1.4.1.4 Reliability

This leads us to reliability, which is a measure of the consistency or accuracy of an outcome measure. Reliability, in contrast to repeatability, is defined relative to between-individual variation. It is defined as the proportion of the total variance which is due to 'true' differences, as opposed to differences arising due to error (itself caused by measurement error or within-individual differences).

$$reliability = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2} = \frac{\sigma_t^2}{\sigma_{tot}^2} \tag{1.8}$$

where the subscripts refer to the sources of variation: t true, e error, and tot total. As described

in equation 1.8 and shown in figure 1.8, it represents the relative proportion of variance in the data attributable to true differences as compared to the combination of true and error. This can be explained by the fact that accuracy of a measurement can only be judged in terms of the question being asked: a common stopwatch can reliably be used to compare the times of runners completing a marathon and group them by percentiles. However, the same stopwatch is not considered sufficiently accurate to decide which medal should be awarded to whom at the 100m sprint in the Olympic games. When true differences between individuals are in the order of hours, then inaccuracies in the order of a few seconds are unimportant. Conceptualised in this way, reliablity can be treated as a measure of *distinguishability* of measurements[93]: the ability to separate individual measurements into those which are relatively high compared to the rest of the group, and relatively low compared to the rest of the group. With poor reliability, the measure itself provides insufficiently accurate information to be able to answer this question. This is metaphorically depicted in the leftmost two panels of the front cover, in which the cats can and cannot be easily distinguished. If measurement error (i.e. within-individual variation) is large, but the between-individual variation is much larger, then individuals can still be meaningfully separated into those who exhibit high outcome values, and those who have low outcome values. Similarly, even if a measure is extremely accurate, it is still incapable of meaningfully distinguishing between individuals who all obtain the same outcome value. As such, reliability is a relevant measure for understanding between-individual comparisons.

There exist several different ways to estimate the reliability of measurements. Within psychometrics (e.g. questionnaires), reliability can also be assessed from the consistency of different item responses within the particular scale: this is referred to as internal consistency and can be assessed using measures such as Cronbach's α. This is not possible for PET and for other measures where the single measurement cannot be broken down into constitutent parts. For these measures, test-retest reliability is a more appropriate measure of reliability.

Test-retest reliability for continuous scales is typically defined using the intraclass correlation coefficient (ICC). In measures for which there are no systematic effects of the measurement order, such as PET test-retest studies in which individuals are each measured twice by the same PET system and for which within-individual changes are expected to be negligible, we use the one-way ANOVA fixed effects model[94]. This is the most conservative formulation of the ICC, and is described by the following equation:

$$ICC = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W} \tag{1.9}$$

where MS_B represents the between subjects mean sum of squares, MS_W represents the within subject mean sum of squares, and k represents the number of observations (usually 2). The ICC is an approximation of the true population reliability: while true reliability can never be negative (equation 1.8) one can obtain negative ICC values if the MS_W is larger than MS_B (equation 1.9), in which case the reliability can be treated as zero.

Measures of reliability are important in order to confirm that the data is capable of distinguishing between individuals to a sufficient degree such that comparisons can be made. There have been several suggestions of standard values by which to judge reliability values[94–98]. As will be argued in in Study V, the guidelines of Portney & Watkins (2015)[98] are most relevant for PET studies, which defined values between 0.5 and 0.75 as poor to moderate, 0.75 to 0.9 as good, and above 0.9 as sufficient for measurements used in clinical diagnosis. These standards are comparatively conservative, but can be considered in light of the fact that an ICC value of 0.5 means that the variance in the data due to true inter-individual variability is equal to the variance due to measurement error: this implies that an individual obtaining a score equal to the mean of the group could have either the highest, or the lowest, underlying true value of the outcome in the group, however the precision with which the outcome is measured is not such that this could be known from a single measurement (see Figure 1.8).

1.4.2 Replicability

Replicability is ultimate standard by which scientific claims can be validated: provided that an effect is real and robust, then any competent researcher should be able to obtain the same result using the same procedures with adequate statistical power[99]. In this way, replicability refers to the degree to which the conclusions made in a scientific study can be replicated in a new sample (or new data) using the same analysis methods. This is described in Figure 1.9, according to which this can be instantiated through computer code, but also refers to similar methods more broadly.

Replications can be divided into direct and conceptual replications. *Direct replication* refers to repeating an experiment in such a way as to correspond as closely as possible to the original study. The purpose of direct replication is therefore to ascertain the consistency by which an effect can be observed. This might involve performing testing using a new sample, or using different software by which to arrive at the results, but doing so in such a way as to adhere to the procedures of the original study. *Conceptual replications*, on the other hand, refer to new experiments to test the predictions of a particular theory. The goal of conceptual replications is to determine the credibility of a theoretical hypothesis[99].

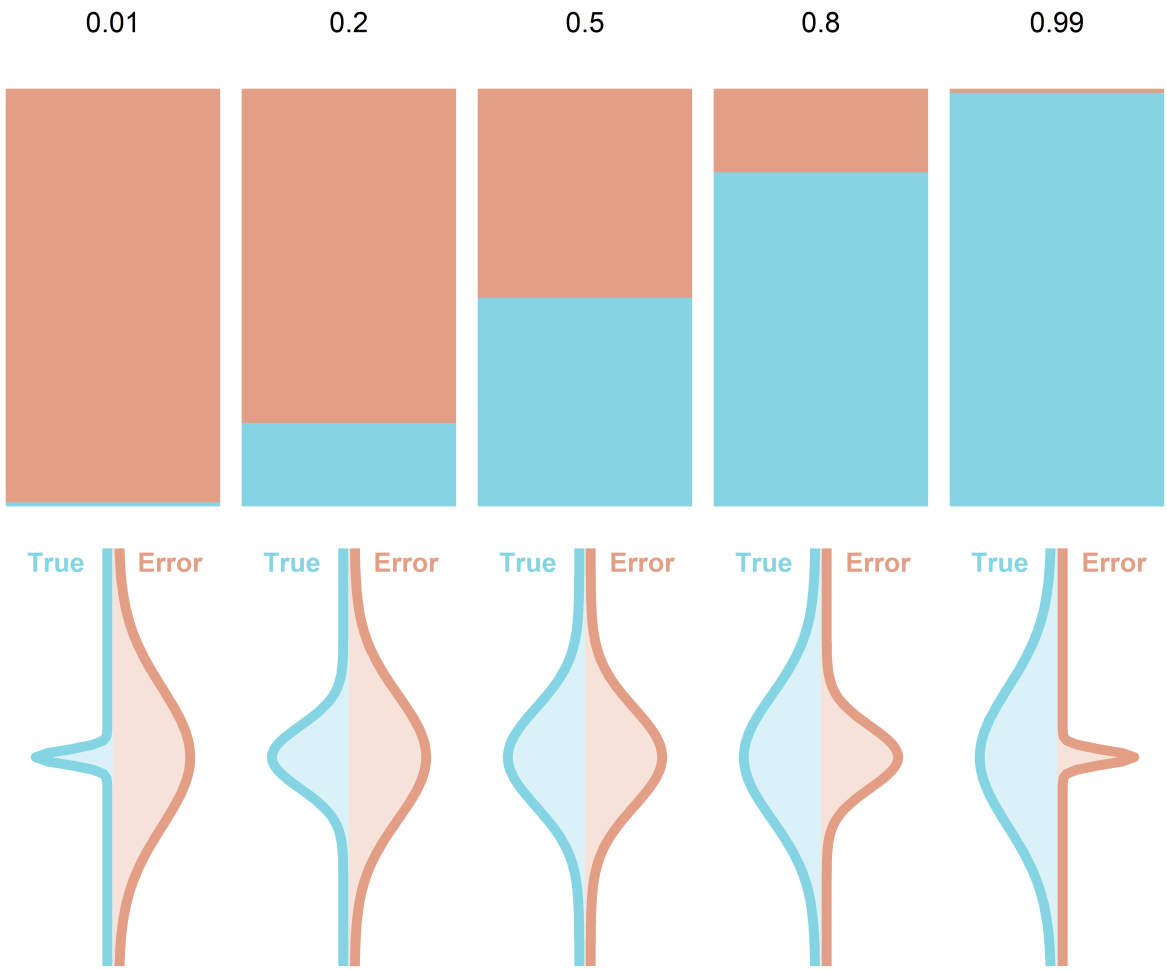

Figure 1.8: Relative variance due to true inter-individual variance in blue (i.e. between-individual variability of the underlying 'true' values), and measurement error in red (i.e. within-individual variability) for different ICC values depicted by their proportional contributions to the measured variance (above) and by their density distributions showing the size of the distributions relative to one another (below).

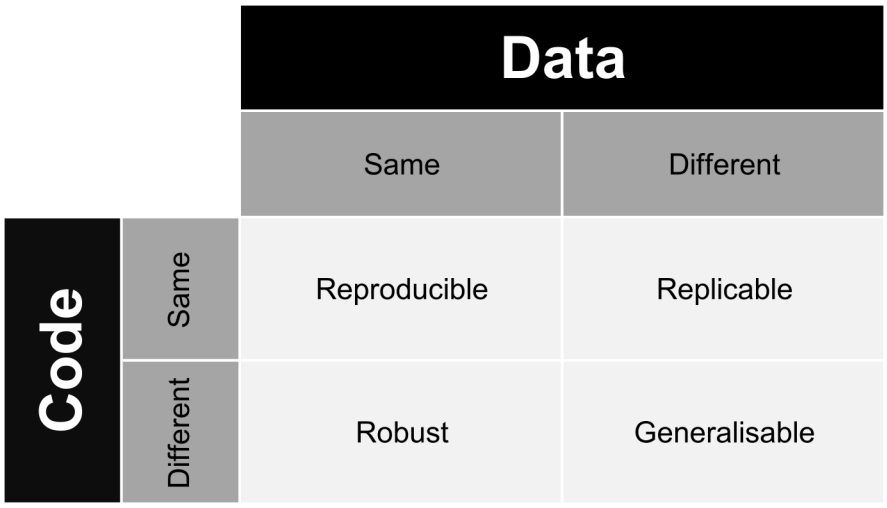

Figure 1.9: Separating a scientific investigation into the constituent data on one hand, and all the operations and analyses conducted on that data on the other, we can describe the ability to come to the same outcome into reproducible, replicable, robust and generalisable. The figure is based on the figure from Whitaker (2018) (100), licensed under CC BY.

Given all the challenges for biomedical research described above, as well as the replicability crisis in psychology and biomedical science[57–62], verification of published results is of great importance for scientific progress, and the gold standard by which the veracity of published results can be assessed is by direct replication[101]. It is this that I will henceforth refer to as replicability. This is metaphorically depicted in the middle two panels of the front cover, in which the cat will, given two samples of the PhD student's study environment, exhibit the same behaviour.

1.4.3 Reproducibility

Neuroimaging, including PET, data are highly complex, and have been growing in complexity and size. Neuroimaging data contain information in different formats from different sources arranged in different ways, and its analysis can take many forms. Collection, storage, analysis and sharing of data and results are therefore highly idiosyncratic between or even within groups, and communication of all steps taken through scientific publications is not always feasible[102–104]. This complicates replication efforts and thereby slows scientific progress. As the size of data sets increases, as well as their complexity, so too do their cost. As such, full replication of scientific claims may not even be feasible for certain research questions due to time and expense[105]: Peng (2011)[101] therefore describes a spectrum of reproducibility, ranging from not reproducible (publication only) through the varying degrees of reproducibility, to full

Figure 1.10: The spectrum of reproducibility based on Peng (2011) (101).

replication as the gold standard (Figure 1.10). According to this definition, reproducibility thereby partially accomplishes the role of a full replication, i.e. increasing the veracity of scientific claims. In this thesis, reproducibility will be defined as varying degrees of this spectrum, but excluding both ends.

This has led to calls for computational reproducibility as a minimum standard for assessment of scientific claims, i.e. that researchers share analysis code and data such that all steps are recorded, allowing an independent researcher to reproduce the results and assess their veracity (Figure 1.9). Reproducibility allows more than the validation of reported results, but also functions to accelerate scientific progress, as novel methods can be readily applied and extended by other researchers using the shared code[101,105,106].

Computational reproducibility also increases the potential for the detection and correction of mistakes, which might otherwise be unrecorded if resulting from manual graphical user interfaces. A prominent example is that of the Duke University Scandal[107,108], where unrecorded user actions during data analysis led to clinical trials in which the wrong, potentially harmful, medications were administered to 109 cancer patients based on incorrect results. This error arose as a result of misalignment of two columns of a spreadsheet. It was only through highly complex reverse-engineering of the results that this could be detected, while a reproducible analysis would have permitted rapid detection and correction of this mistake. Another notable example was an economics paper claiming to demonstrate the effectiveness of austerity measures in fiscal policy, which was widely cited in political debates[109]: "surely the most influential economic analysis of recent years"[110]. The original study was not accompanied by the data and code. Upon request, the spreadsheet was later made available, in which it was detected that omissions, questionable statistical procedures and an error in a Microsoft Excel formula led to a complete reversal of the reported effect[111]. It is therefore of great importance that analysis code, and data if possible, are shared, and that tools and methods are developed which allow for all steps to be reported transparently.

For this reason, in the later papers presented in this thesis, we have published the analysis code, as well as the data where possible. For those papers for which the data is shared, this allows reproduction of our results, and for other researchers to be able to experiment with our

procedures both to fully understand them, as well as to apply or extend them for their own purposes. For those papers for which the analysis code is shared, but not the data, this will at the very least allow other researchers to see exactly how we have performed every part of the presented analyses, and to use our same procedures. This is metaphorically depicted in the rightmost two panels of the front cover, in which scientific results are depicted by the cat, and all of the steps between the data (the ink pool) and the final results are visible and recorded.

1.5 Clinical Applications

The principles of reliability, replicability and reproducibility are applied in this thesis to several research questions. Below, I will provide a brief background to each of these questions.

1.5.1 Schizophrenia

Schizophrenia has a lifetime prevalence of 0.7% in the general population[112] and has a very high degree (~80%) of heritability[113]. The disorder gives rise to great adversity both for sufferers of the condition as well as their caregivers. Current pharmacological treatment is unsatisfactory in terms of efficacy and side effects, and there is therefore a great need for improved diagnostic and treatment strategies directed towards the pathological mechanisms of the disease. In particular, finding both behavioural and biological markers to aid diagnosis at a very early phase of the disorder would enable the development of preventative treatment approaches.

The most characteristic symptoms of schizophrenia are hallucinations and delusions. Hallucinations are false perceptions, such as hearing voices. Delusions are irrational or strange beliefs, such as believing that others can hear one's thoughts. Theories about schizophrenia have traditionally considered these perceptual alterations and the formation of irrational beliefs as separate dysfunctions, but recent models of this disorder suggest that these experiences may be attributable to the same core deficit beginning at some of the earliest levels of perceptual processing[114].

1.5.2 Dopamine, the D1 Receptor and schizophrenia

Schizophrenia has been directly associated with dopamine for over fifty years, due primarily to the the fact that antipsychotic drugs block dopamine receptors, that those which bind to dopamine receptors with greater affinity were more clinically effective[115], and that the

extent of occupancy was related to drugs' clinical effectiveness[6,7]. The dopamine hypothesis of schizophrenia has evolved over these years through several iterations[116]. In its first iterations, it was focused on excessive dopaminergic transmission at dopamine receptors. In its second iteration, it came to regard the dopaminergic disturbances as being regionally specific: more specifically striatal hyperdopaminergia and frontal hypodopaminergia. The most recent versions of this theory posit that psychosis in schizophrenia is associated with elevations of presynaptic striatal dopamine function as a 'final common pathway': there exist many causes of schizophrenia., all of which may interact, but they all converge in striatal presynaptic dopamine hyperfunction. These alterations in dopamine function are associated with psychosis, rather than schizophrenia in general. The effects of this dysfunction of the dopamine system influence evaluation of perceived stimuli through aberrant salience attribution[117]. This theory therefore differs from previous conceptions primarily in its focus on presynaptic changes in dopamine function, and in downplaying prefrontal hypodopaminergia.[116]. However, there has recently been some *in vivo* evidence in support of cortical hypodopaminergia too[118].

There has been a great deal of *in vivo* evidence, shown in numerous PET studies, for elevations in presynaptic dopamine function in schizophrenia[116]. These studies have focused on dopamine release following amphetamine challenge as well as presynaptic dopamine synthesis. There has also been evidence of elevations of presynaptic dopaminergic synthesis in individuals at high risk of developing schizophrenia, of a smaller magnitude compared to those individuals with schizophrenia[119]. Furthermore, presynaptic dopamine synthesis was shown to be higher at baseline in those who subsequently transitioned to schizophrenia compared to those who did not[120], and increased further in these individuals following the transition to schizophrenia[121].

In contrast, only a few PET studies have examined the D1 receptor (D1R) in schizophrenia. Compared with the D2R, there is a much higher concentration of D1R in the cortex[122], and the frontal, and especially the dorsolateral prefrontal, cortex is thought to be a crucial brain region for understanding the biological basis for schizophrenia[123–125]. In-vivo studies of the D1R in schizophrenia patients have yielded mixed results (Table 1.1). Initial studies found lower[126], higher[127], or no difference[128] in the availability of D1R in the frontal cortex compared to healthy control subjects. This led to concerns as to whether there were systematic differences between what was detected with the two radioligands employed in these studies ([^{11}C]SCH23390 or [^{11}C]NNC112). The former two research groups have both subsequently replicated their own respective results, in a sample of chronic medicated patients[129], and in a subsample of drug naive patients[130] respectively. Further, both groups examined patients and controls using both radioligands[129,131], leading both to conclude that any differences were unlikely to be a result of differences between the two tracers. In another small sample of twin pairs discordant for schizophrenia, Hirvonen et al. (2006)[132] observed decreases in D1R binding

Table 1.1: PET studies comparing D1-R binding in patients with schizophrenia or schizophreniform psychosis to that of healthy control subjects

Publication	Subjects SCZ(DN)/HC	Radioligand	Differences
Okubo et al. 1997	17(10)/18	SCH	↓: PFC
Abi-Dargham et al. 2002	16(7)/16	NNC	↑: DLPFC
Karlsson et al. 2002	10(10)/10	SCH	no sign. diff.
Hirvonen et al. 2006	9(0)/11	SCH	↓: CAU,PUT,CX
Kosaka et al. 2010	6(0)/12	both	↓: FC,ACC, TC,STR
Abi-Dargham et al. 2012	25(12) / 48	NNC	↑ (DN): DLPFC, MPFC,OFC
Poels et al. 2013	7(4)/11	SCH	no sign. diff.
Study IX	18(18)/17	SCH	↓: DLPFC

SCZ=patients with schizophrenia or schizophreniform psychosis; DN=Drug Naive; HC=healthy control subjects; NNC=[^{11}C]NNC112; SCH=[^{11}C]SCH23390; PFC=prefrontal cortex; DLPFC= dorsolateral prefrontal cortex; CAU=caudate nucleus; PUT= putamen; CX=cortical regions; FC= frontal cortex; ACC=anterior cingulate cortex; TC=temporal cortex; STR=striatum; MPFC=medial prefrontal cortex; OFC=orbitofrontal cortex.

in chronic, medicated schizophrenia probands compared to controls. In contrast, higher levels were shown in monozygotic unaffected co-twins, i.e. those individuals at high genetic risk.

Of note, in studies where both drug naive and either medicated or drug free patients were examined, the former group has shown numerically lower D1R binding[126,127,130,131]. This may be explained by a reduction in D1R due to antipsychotic treatment as has been shown in experimental studies of non-human primates (NHPs)[133,134] (although see Knable et al. 1996[135]), or in the case of ongoing medication, direct D1-R occupancy[7,132]. To avoid this confounding factor, future investigations of differences in the availability of the D1R in schizophrenia need to focus on the early stages of the illness, before antipsychotic treatment, or changes prior to its onset which lead certain individuals to be more prone to psychosis than others.

Though all of the PET studies of the D1R in schizophrenia patients have been conducted with small sample sizes, and therefore with low statistical power, a tentative interpretation of the results is that drug-naive patients with psychosis disorders, and potentially also unmedicated individuals at high genetic risk for schizophrenia, show higher D1R binding in frontal cortex[136].

1.5.3 Immune activation and TSPO in schizophrenia

In addition to the dopamine system, the association of schizophrenia with immune function also has a long history starting as early as the 1930s[137]. Findings of elevated schizophrenia risk for people born in the winter months[138,139] led to suggestions that the disease may be a result of immunological disturbances following prenatal viral exposure. Evidence for a role for the immune system in the pathophysiology of schizophrenia has since accumulated from the fields of genetics, epidemiology and immunological research[140–142]. Although the involvement of the dopamine system in schizophrenia has been clearly demonstrated, the underlying mechanism by which dopaminergic dysfunction might occur has been suggested to possibly lie in disturbances of the glutamatergic system caused by immune function[143,144]. In this case, immunological changes may precipitate the dopamine dysfunction thought to represent the "final common pathway" for schizophrenia[116].

Translocator protein (TSPO) is expressed throughout the body and brain[145], but importantly is expressed in glial cells, including microglia within the brain. Active brain disease causes the activation of microglia, and this change in state is associated with *de novo* expression of TSPO[146,147]. For this reason, quantification of TSPO within the brain is thought to provide an index of microglial activation, or neuroinflammation[148,149]. PET imaging using tracers which bind to TSPO is currently the most established measure of *in vivo* neuroinflammation[150].

PET tracers for quantification of TSPO are divided into first-generation (i.e. [^{11}C]-(R)-PK11195) and second-generation tracers. [^{11}C]-(R)-PK11195 exhibits a low signal-to-noise ratio[151,152] due to its high lipophilicity and hence high nonspecific binding, as well as poor test-retest reliability[153,154], however it remains the most commonly used PET TSPO tracer. Concerns regarding the low signal-to-noise ratio of this tracer led to the development of second-generation TSPO tracers, including [^{11}C]PBR28, [^{11}C]DPA-713, [^{11}C]ER176 and [^{18}F]FEPPA among others, which exhibit higher signal to background (i.e. BP_{ND}) ratios[151,152,155], as well as higher reliability[156–159].

There have been numerous studies attempting to measure *in vivo* neuroinflammation in schizophrenia using PET. The first studies made use of the [^{11}C]PK11195[160,161], indicating increases in patients compared to controls. Following the introduction of second-generation TSPO radiotracers, there was a diversification of results. Later studies using [^{11}C]PK11195 with larger samples showed no differences[162,163], or higher levels only in medicated patients with no differences observed between healthy controls and antipsychotic-free patients[164]. Studies using second-generation TSPO tracers, including [^{11}C]DAA1106, [^{11}C]PBR28, [^{11}C]DPA-713 and [^{18}F]FEPPA, observed no differences[165–168], increases[169] as well as decreases[170].

1.5.4 Delusional Ideation and Self-Transcendence

As described earlier, hallucinations and delusions are central symptoms of schizophrenia. While psychotic patients have more of these experiences than healthy populations do, there is a wide distribution of the number of delusional beliefs[171,172] and anomalous perceptions[173] held by healthy populations. These experiences in healthy populations exhibit various properties which suggest that they may be related to the same experiences in patients: delusional beliefs and anomalous perceptions appear to co-vary to a large extent[173], there are more of these beliefs held by relatives of schizophrenia patients[174], and these subclinical symptoms may constitute an important risk factor for later psychosis[175,176]. This provides evidence both for the validity, and of the relevance, of these behavioural dimensions for understanding psychosis itself.

Just because symptoms associated with psychosis are continuously distributed within the general population, this does not necessarily imply that psychosis itself is not qualitatively different from normal experience[177]. This necessitates careful definition of what exactly is being referred to, and what exactly it is that lies along the continuum. It has been suggested that psychotic experiences should be separated from true subclinical psychotic symptoms (associated with distress and help-seeking), but even psychotic experiences have an estimated prevalence of less than 10%[176,178]: these symptoms are still described as being 'clinically relevant'[178].

The types of beliefs and experiences measured by the PDI[171,172] and CAPS[173] scales respectively, are comparatively much more common, and for the most part are primarily not of clinical relevance unless present in large numbers or associated with significant distress. Rather, these scales are considered an index of delusional ideation and anomalous perceptions as manifestations of a latent schizotypy which may predispose one to developing psychosis. These measures do not represent gradations of disease, and high scores do not deterministically produce the disease. Rather, an individual with a very high score on these scales may never develop psychosis, but may have been more 'prone' to develop psychosis than another individual who, through environmental exposure, *does* go on to develop psychosis. While there exists considerable disagreement about what schizotypy truly represents[179], this view is consistent with a version of schizotypy which accepts a continuum among the general population, but which posits that development of a psychotic episode itself is not attributable simply to an extreme number or strength of this trait in and of itself.

As such, these scales are thought to be of utility for understanding the continuum of psychological and/or biological processes which may underlie these symptoms in psychosis, and have been shown to be related to various other behavioural models of psychotic

symptoms[180–182].

The Temperament and Character Inventory (TCI) is a model of the structure and development of personality, and one of its subscales is the Self-Transcendence scale[183]. This scale was created to refer to the degree to which an individual feels part of nature and the universe at large, as well as extraordinary experiences: it is a measure of creativity and spirituality[183,184]. These extraordinary experiences, however, include experiences such as exta-sensory perception, which bears more than a passing resemblance to items from the PDI questionnaire. Indeed, it was later found that scores on this scale were in fact associated with proneness to developing psychosis as well as psychotic phenotypes[185–190]. This can be taken both to lend more evidence for this theory of delusional ideation as psychosis proneness, but also to suggest that this scale may, in addition to the PDI, be of relevance for understanding this trait.

Chapter 2

Aims

The central aims of this thesis can be conceptualised in terms of proximal and distal aims.

Proximal aims

The proximal aims of this thesis are focused on advancing the ability of PET research to answer applied clinical research questions in a manner which is as robust as possible as follows:

I **Reliability**: Measure and evaluate the reliability of outcome measures for answering clinical research questions, and determine the influence of various image analytic and methodological factors on the reliability of these measures.

II **Replicability**: Evaluate the correctness of published and preliminary scientific results by replication.

III **Reproducibility**: Conduct research in as robust and transparent a manner as possible to allow others to better evaluate claims, reduce the potential for errors, and increase the ease and speed of transmission of methods and tools.

Distal aims

The distal aim of this thesis is the study and characterisation of the disturbances underlying, and leading to the development of, schizophrenia. We examine the following:

I **Dopamine system**: Evaluate the hypothesis for changes in the brain dopamine system, and the D1 receptor in particular, associated with schizophrenia and proneness for

developing schizophrenia, using PET.

II **Immune system**: Evaluate the hypothesis that schizophrenia is associated with upregulation of the immune function in the brain, using PET.

Chapter 3

Materials and Methods

This chapter provides a general description of the methods used throughout the following thesis.

3.1 Participants

All studies involving collection of PET data from human participants were approved by the Regional Ethics and Radiation Safety Committee of the Karolinska Hospital, and all subjects included in the studies provided written informed consent prior to their participation. The thesis also includes data which was collected in other PET centres in the United Kingdom, the United States of America and in Canada, however these data consisted only of outcome measures from already-published studies to answer the same research question as had originally been posed, and thereby did not require additional ethical approval or consent.

Studies I and II: The participants included in these studies consisted of sixteen healthy control subjects, each of whom were examined twice with [^{11}C]SCH23390. All participants underwent a screening procedure and were deemed to be healthy with no history of significant psychiatric or somatic illness. Participants were aged between 21.8 and 35.0 years, and all were male.

Study III: The participants included in this study consisted of twelve healthy control subjects, each of whom were examined twice with [^{11}C]PBR28. The study consisted of six medium-affinity binders (MABs) and six high-affinity binders (HABs)[191–193]. The mean age of participants was 23.9 (SD 2.99), and the sample consisted of six males and six females. These individuals are the same as those included in a previous test-retest study with this tracer[158].

Study IV: The participants included in this study consisted of healthy control subjects, of whom 56 individuals aged between 22.0 and 55.4 years old were examined with [^{11}C]WAY-100635, and 40 individuals aged between 21.9 and 55.4 years old were measured using [^{11}C]MADAM. These individuals consisted of healthy male control subjects from previous studies conducted within the group using the same PET system and the same measurement and reconstruction protocols[194–201].

Study V: This study did not make use of any new participants, and instead consisted of an analysis of summary statistics from several previously published studies[20,158,202,203].

Study VI: The participants included in this study are the same as those included in Study IV who underwent PET measurements using [^{11}C]WAY-100635 who had also completed the Temperament and Character Inventory (TCI) scale. The final sample consisted of 50 healthy male participants in the same age range as for Study IV.

Study VII: This study was a meta-analysis of published findings using individual participant data (IPD). Research groups who had published studies comparing TSPO levels in schizophrenia patients with controls using arterial blood sampling and using second-generation TSPO PET tracers, were asked to provide V_T values quantified using the two-tissue compartment model[166–170]. The final sample consisted of 77 healthy controls (35 female, 56 HABs) and 75 patients with psychosis or schizophrenia (24 female, 52 HABs) aged 33.9 $\pm$ 12.6 and 35.4 $\pm$ 15.1 (mean $\pm$ SD) respectively.

Study VIII: This study consisted of four substudies conducted using data from three different cohorts of subjects. A first cohort consisted of 132 subjects (72 female) aged between 22 and 76 who had completed psychometric questionnaires. The second cohort consisted of a cohort of individuals who were each measured once with PET and [^{11}C]SCH23390. Of these individuals, 27 (8 female), aged between 23 and 76, completed psychometric questionnaires at the time of PET. Of the same cohort, those individuals who were male and between the ages of 20 and 35 at the time of PET, were contacted by letter and asked to complete an additional psychometric questionnaire online. This sample consisted of 41 participants, who completed the questionnaires between 4.8 and 12.7 years after their PET studies. An third cohort was collected consisting of 20 healthy male participants aged between 22 and 35, and each measured at least once with [^{11}C]SCH23390 and who completed psychometric questionnaires.

Study IX: The participants included in this study consisted of 17 healthy control subjects and 18 first-episode drug-naive psychosis patients, each of whom were measured with PET and [^{11}C]SCH23390. Participants were aged between 22 and 52 years.

3.2 Magnetic Resonance Imaging Procedures

All participants in all studies underwent MRI measurements to describe the underlying brain anatomy. In all studies but Study IX, T1-weighted MRI images were acquired for all individuals for delineation of anatomical regions of interest. T2-weighted images were acquired for all individuals who underwent PET at Karolinska Institutet and examined for structural pathology. In Study IX, only T2-weighted images were acquired for some individuals, which were used for anatomical delineation.

Three different MRI systems were used for PET measurements conducted at the Karolinska Hospital. These include the 1.5T Siemens Magnetom Avanto system (Erlangen, Germany) (Studies I, II, XIII), the 1.5T GE Signa system (Milwaukee, WI) (Studies IV, VI, IX), and the 3-T General Electric Discovery MR750 system (GE, Milwaukee, WI) (Study III).

3.2.1 Definition of regions of interest

In the thesis, regions of interest (ROIs) were defined using manual, automated and semi-automated methods. Manual ROI delineation was performed using in-house software[204] in which ROIs are drawn on subsequent slices of the anatomical MR image. Our primary method for automated delineation of ROIs was FreeSurfer (5.0.0, http://surfer.nmr.mgh.harvard.edu/)[205–208], which defines ROIs in the space of each individual MR. Our secondary method for automated ROI delineation was the use of volumetric normalisation methods. This involves normalising each individual brain to MNI (Montreal Neurological Institute) space, and saving the warping parameters. Subsquently, these warping parameters can be inverted and used to warp ROIs defined in MNI spatial coordinates back to individual space. For the reference region in particular, we aimed not to have the most anatomically correct region, but rather the specific areas of the anatomical reference region (in this case the cerebellum) with the best properties for quantification. We therefore made use of a customisation of the defined region which would exclude regions of the reference region which were in close proximity to either the CSF or to higher-binding regions. Semi-automated methods for ROI delineation involved the approximate delineation of ROIs, and using the image itself to refine the delineation by selecting the voxels with the highest intensity[209].

3.3 Positron Emission Tomography Procedures

3.3.1 Radioligands and Targets

[^{11}C]SCH23390[210] was used in studies I, II, VIII and VI to measure the dopamine D1 receptor. This receptor is highly concentrated in the striatum, lower in cortical regions, and negligible in the cerebellum[122]. [^{11}C]SCH23390 binds primarily to the dopamine D1 receptor, although 5-HT$_{2A}$ receptor binding represents a non-negligible proportion of the specific binding (estimated to be approximately a quarter) in cortical regions[211].

[^{11}C]PBR28[212] was used in studies III and VII. This is a second-generation radiotracer for the 18kDa translocator protein (TSPO). This target was initially described as a peripheral benzodiazepine receptor, but it has since been shown to be expressed throughout the body and brain[145]. Importantly, it is expressed in glial cells, including microglia. Active brain disease causes the activation of microglia, and this change in state has been associated with *de novo* expression of TSPO[146,147]. For this reason, PET tracers binding to TSPO are used as an estimate of microglial activation, or neuroinflammation[150]. This radiotracer binds throughout the brain, meaning that there is no suitable reference region[213]. Comparison of individuals assessed using this tracer is complicated by the fact that the binding affinity is dependent on a polymorphism in the gene encoding the protein[191–193] which is not thought to affect the biological function of TSPO. However, this results in differences in binding estimates: individuals are thus described as high-affinity binders (HABs, high-affinity homozygotes), medium-affinity binders (MABs, heterozygotes) and low-affinity binders (LABs, low-affinity homozygotes). These differences must be accounted for during statistical analysis.

[^{11}C]WAY-100635[214] was used in studies IV and VI. This is a radiotracer for the serotonin 1A (5-HT$_{1A}$) receptor, which is highly concentrated in the hippocampus and neocortex, and which is expressed in very low quantities within the basal nuclei. Binding in the cerebellum is low in most individuals, but exceptions have been noted, thereby complicating its use as a reference region for quantification[215,216].

[^{11}C]MADAM[217] is a radiotracer for the serotonin transporter (5-HTT), and was used in Study IV. This tracer exhibits high binding in the striatum (especially the putamen), thalamus and brainstem, moderate binding in the cingulate cortex and limbic cortex, low binding in the neocortex, and very low to negligible binding in the cerebellar cortex[196].

3.3.2 Image Acquisition and Reconstruction

PET measurements which took place at the Karolinska Hospital (i.e. excluding Study VII) were performed using two PET systems, namely the ECAT Exact HR 47 (CTI/Siemens, Knoxville, TN), and the High Resolution Research Tomograph (HRRT) (Siemens Molecular Imaging). These will henceforth be referred to as the HR and HRRT PET systems.

For PET images acquired using the HR, spatial resolution ranges from 3.6 mm full width half maximum (FWHM) at the centre of the field to 4.5 mm tangentially and 7.4 mm radially at 20 cm from the centre[218]. Prior to each PET measurement, transmission scans were performed using three rotating ^{68}Ge rods in order to correct for signal attenuation. For all studies except for Study IX, the camera was run in 3D mode. For Study IX, acquisition was performed in 2D for most measurements. Data were reconstructed using filtered backprojection using a Hann filter with a 2mm cutoff frequency.

For PET images acquired using the HRRT, spatial resolution ranges between 2.5mm FWHM at the centre of the field, to 3.5 mm at 14 cm from the centre[219]. Prior to each PET measurement, transmission scans were performed using a single ^{137}Cs source. Measurements were acquired in 3D. Reconstruction was performed using the 3D-OP-OSEM reconstruction with 16 subsets and 10 iterations.

3.3.3 Arterial Blood Sampling

Arterial blood sampling was performed for the PET examinations in Study III. During the first minutes of the acquisition, blood samples were acquired continuously using an automated blood sampling system (Allogg AB, Sweden). Manual blood samples of 1-3 ml were also acquired throughout the acquisition. Radioactivity was measured using a well counter cross-calibrated with the PET system. Plasma samples were obtained following centrifugation of the blood. Plasma radioactivity was measured using the same well counter, from which plasma-to-blood ratios could be derived. The plasma parent fraction, i.e. the fraction of the unmetabolised tracer compound in the plasma, was assessed using high performance liquid chromatography.

For obtaining metabolite-corrected arterial plasma input function curves, we first performed dispersion correction on the blood measurements obtained using the automated blood sampling system, and then derived blood curves using linear interpolation of the blood radioactivity using both automated as well as manual samples. Linear interpolation was applied to measured plasma-to-blood ratios, and this ratio was multiplied by the blood curve to obtain an estimate of plasma radioactivity. Linear interpolation was applied to measurements of the plasma parent

fraction, which were multiplied by the plasma concentrations to obtain metabolite-corrected arterial plasma input function curves.

3.3.4 Kinetic Modelling and Quantification

In the majority of the quantification performed within this thesis, we made use of the simplified reference tissue model (SRTM)[42]. This model makes use of nonlinear least squares estimation to derive estimates of three parameters: R_1 (the relative target to reference rate of delivery, i.e. $\frac{K_1}{K_1'}$), k_2 (the efflux rate constant), and BP_{ND}. This model is based on four assumptions, namely that 1. the reference region has no specific binding, 2. the kinetic behaviour of the tracer can be represented by a one tissue compartment model (1TCM) in both the target and reference regions, 3. the blood volume for both the target and reference regions is negligible, and 4. that both the target and reference regions have the same V_{ND}. SRTM is described by the following equation.

$$C_T(t) = R_1 C_R(t) + (k_2 - \frac{R_1 k_2}{1 + BP_{ND}^{target}}) C_R(t) \otimes e^{-\frac{k_2}{1+BP_{ND}^{target}}t}$$

In studies for which parametric imaging was used, we made use of the 3D stationary wavelet-transform aided parametric imaging (WAPI)[220,221] employing the non-invasive Logan plot fitted with a multilinear regression[222]. This model is fitted using ordinary least squares, and requires a fixed value of k'$_2$ (which we define as 0.1 by default), as well as a specification of the t* (the point in time at which the fitted line becomes linear). It yields two parameters, namely BP_{ND} and b, the latter of which can be discarded. This model is described by the following equation.

$$\int_0^t C_T(\tau)d\tau = DVR\left(\int_0^t C_R(\tau)d\tau + \frac{C_R(t)}{k_2'}\right) + b\, C_T(t)$$

We made use of the two-tissue compartment model (2TCM) using metabolite-corrected arterial plasma (AIF) for quantification of [^{11}C]PBR28. This model makes use of nonlinear least squares estimation to derive estimates of the rate constants K_1, k_2, k_3 and k_4, from which V_T is calculated using the following equation: $V_T = \frac{K_1}{k_2}(1 + \frac{k_3}{k_4})$.

We also made use of several semi-quantitative estimates of uptake. These included the standardised uptake value, described as $SUV(t) = \frac{C_{img}(t)}{IR/BW}$, where C_{img} is the radioactivity concentration in the image for a particular time frame, IR is the injected dose of radioactivity,

and BW is the bodyweight. This measure is therefore an index of the tracer uptake relative to the injected dose per unit body mass. This measure is usually calculated as the AUC for a particular time interval. We also employed ratio methods: the SUV ratio (SUVR) and the distribution volume ratio (DVR), which are ratios of SUV and V_T values in the target compared to a reference region.

3.4 Questionnaires

In studies VI and VIII, we make use of questionnaires. In Study VIII, we made use of the 21-item version of Peters Delusion Inventory (PDI)[171,172]. In both studies, we made use of the Temperament and Character Inventory (TCI)[183], translated into Swedish[223]. In Study VI, we made use of the self-transcendence scale and its spiritual acceptance subscale. In Study VIII, we made use of the same scale, but selected only a subset of items to define an instrument for the measurement of delusional ideation specifically, rather than self-transcendence. We performed psychometric validation of these items, and derived a novel scale for this construct which showed acceptable reliability and convergent validity.

3.5 Statistical Analysis

In studies II, IV, VI, VIII and IX, we made use of various classic forms of linear regression including correlations, multiple linear regression, t-tests and two-way ANOVA, using p values to draw inferences. Additionally in Study I, we made use of non-parametric Mann-Whitney U Tests since normality of residuals could not be assumed. In the exploratory analysis of Study XIII, we performed a multiverse analysis[224], since we could not define a favoured analysis method a priori. This means that the results were transparently presented following each potential analysis decision, i.e. along each path of the "garden of forking paths".

We also used principal components analysis (PCA) in Study III for dimension reduction in order to examine the correlational structure of the data. Distributions were represented using histograms, kernel density estimation displayed in density plots and violin plots. Two-dimensional kernel density estimations were also represented by colour in scatter plots in Study I.

3.5.1 Bayesian Analysis

We used Bayesian methods to perform both hypothesis testing (Studies VI-IX) as well as parameter estimation (Study VIII). Bayesian model fitting was performed both using analytical solutions using *JASP*[225], as well as using Markov Chain Monte Carlo (MCMC) using the Gibbs sampler *JAGS*[226], as well as the Hamiltonian Monte Carlo sampler *STAN*[227] (the latter performed using the *brms* interface[228]) called from R[229]. In studies making use of Bayesian hypothesis testing, we made use of Bayes Factors (BFs), comparing the average likelihood of the alternative hypothesis to the (nested) null hypothesis, using the Savage-Dickey ratio[230].

We made use of both regularising and informative priors. Regularising priors are centred at zero and decrease in probability with more extreme values. We made use of normal distributions, for which the standard deviation is equal to the size of the tested effect: this means that 68% of the prior probability density is allocated to effect sizes smaller than the tested effect size, and only 5% of the probability density to effect sizes more than twice as large, based on the recommendation of Dienes (2014)[231].

Informative priors were defined for accounting for the effects of confounders, and specifically the effects of age. In most studies in this thesis, age is expected to have a a non-negligible effect on various outcome parameters. However, in the examined data, neither the age ranges nor the sample sizes are large, which might have allowed us to accurately assess the effects of age. Rather, we can see that there appears to be an effect, but deriving an accurate estimation of its magnitude based on our data is not possible. The usual solution would be to add age to the regression model and hope that it is relatively accurately estimated. However, using Bayesian modelling, we can define an informative prior over the expected degree to which age is expected to affect the measured values, and then update this estimate based on our data. This both corrects for the effects of age, as well as constrains this correction to reasonable values of the relationship based on previous studies using larger samples and wider age ranges.

Informative priors were also used in this thesis in the studies which involved a replication, or more specially, made use of replication BFs[232–234]. These tests evaluate the success of a replication attempt by defining the results of an initial study as a posterior distribution. This posterior distribution is utilised as a prior for the parameter of interest in the replication attempt, and the BF is calculated using a Savage-Dickey ratio. In so doing, the replication BF compares a skeptic's null hypothesis with the proponent's initial results, including their uncertainty. The replication BF can broadly be interpreted as a measure of how much support there is in data for a successful replication relative to a failed replication. Alternatively, it can be formulated as the relative likelihood of the new data originating under each of the two

hypotheses, or as the relative change in the likelihood of the estimated effect being equal to 0 before, and after, having observed the data from the replication study.

Linear mixed effects modelling (multilevel modelling) was employed in the meta-analysis. Such models allow for the specification of a hierarchical structure of data to exploit similarities between data belonging to different clusters, thereby leading to improved estimates and predictions[235]. Multilevel modelling has been applied in numerous fields of research, and has been argued to be a more sensible default method by which statistical inference should be performed in general[236].

Chapter 4

Results and Discussion

In this section, I will outline the relevant results and conclusions from each study, describe how the themes are represented by each study, and follow this with a transparency statement for certain studies.

4.1 Study I: Reliability of volumetric and surface-based normalisation and smoothing techniques

In neuroimaging, the conventional approach for parametric analysis is to apply normalisation and smoothing procedures in three dimensions. However, due to the high surface-area-to-volume ratio in the cortex, volumetric (three-dimensional) methods entail that values originating in voxels belonging to different tissue types are averaged together. Surface-based (two-dimensional) methods were proposed as an alternative approach which would be expected to minimise this issue for cortical regions[39], and it was subsequently shown that this method greatly reduced intersubject variance and bias in PET data acquired using an HRRT PET system using [^{11}C]SB207145[40].

In this study, we aimed to extend these findings to determine whether they were also reflective of improved test-retest repeatability and reliability. In the previous study, kinetic modelling was always performed after smoothing, however our group has shown excellent performance of a wavelet-based method for noise reduction (wavelet-transform aided parametric imaging, WAPI)[220,221] which performs minimal smoothing while enhancing signal-to-noise ratio. We therefore also aimed to explore whether kinetic modelling performed before or after smoothing could rescue the performance of volumetric methods, or further improve the performance of

surface-based methods.

We found that surface-based methods produced higher BP_{ND} values, with less dispersion (coefficient of variation, COV) and with less bias. For the comparison of test and retest outcomes, we showed that surface-based methods improved repeatability, and decreased measurement error (SEM). However, we found that surface-based methods exhibited lower reliability (Figure 4.1).

Due to the fact that reliability values appeared to show an advantage for volumetric methods in contrast to all other tested metrics, we examined this outcome in more detail. By comparing the reliability (ICC) with other outcomes at the level of individual voxels, we discovered that those voxels for which the ICC was highest for volumetric methods, were also the voxels with poor repeatability, high bias and negative BP_{ND} (Figure 4.2). Of the voxels with ICC>0.75 using an 8mm volumetric smoothing kernel, 44% had an APD>90%, and 42% had BP_{ND}<0. Using the comparable surface-based smoothing kernel using dynamic data, only 13% had an APD>90%, and 1.3% had BP_{ND}<0. This means that the apparent improvement in reliablity for volumetric methods is likely to be artefactual. Further, this means that volumetric normalisation and smoothing methods are inducing systematic differences between individuals, presumably as a result of subtle differences in anatomy.

We additionally showed that the use of WAPI was not sufficient to cause volumetric methods to exhibit similar performance to surface-based methods. Further, for surface-based methods we showed that with a small degree of smoothing, modelling BP_{ND} using WAPI prior to smoothing caused improvements over modelling BP_{ND} after smoothing, but that this improvement disappeared and even reversed with larger smoothing kernels, suggesting that the less computationally-demanding surface-based smoothing method can be used for noise reduction instead of the more computationally expensive WAPI method, provided that the degree of smoothing is sufficient.

In conclusion, we show that surface-based methods for normalisation and smoothing appear not only to improve the spread and result in higher BP_{ND} values as previously reported, but also improve test-retest performance. Although reliability appeared to be increased for volumetric methods, we showed that this reflected systematic bias in BP_{ND}. This bias, due to its non-stochastic nature, is likely to increase the possibility of false positives in applied studies, especially in studies of patient groups who are known to exhibit anatomical differences. These improvements in performance using surface-based methods were calculated to correspond with the need for a sample size of approximately half the size to detect a difference between groups of a given magnitude compared to using volumetric methods.

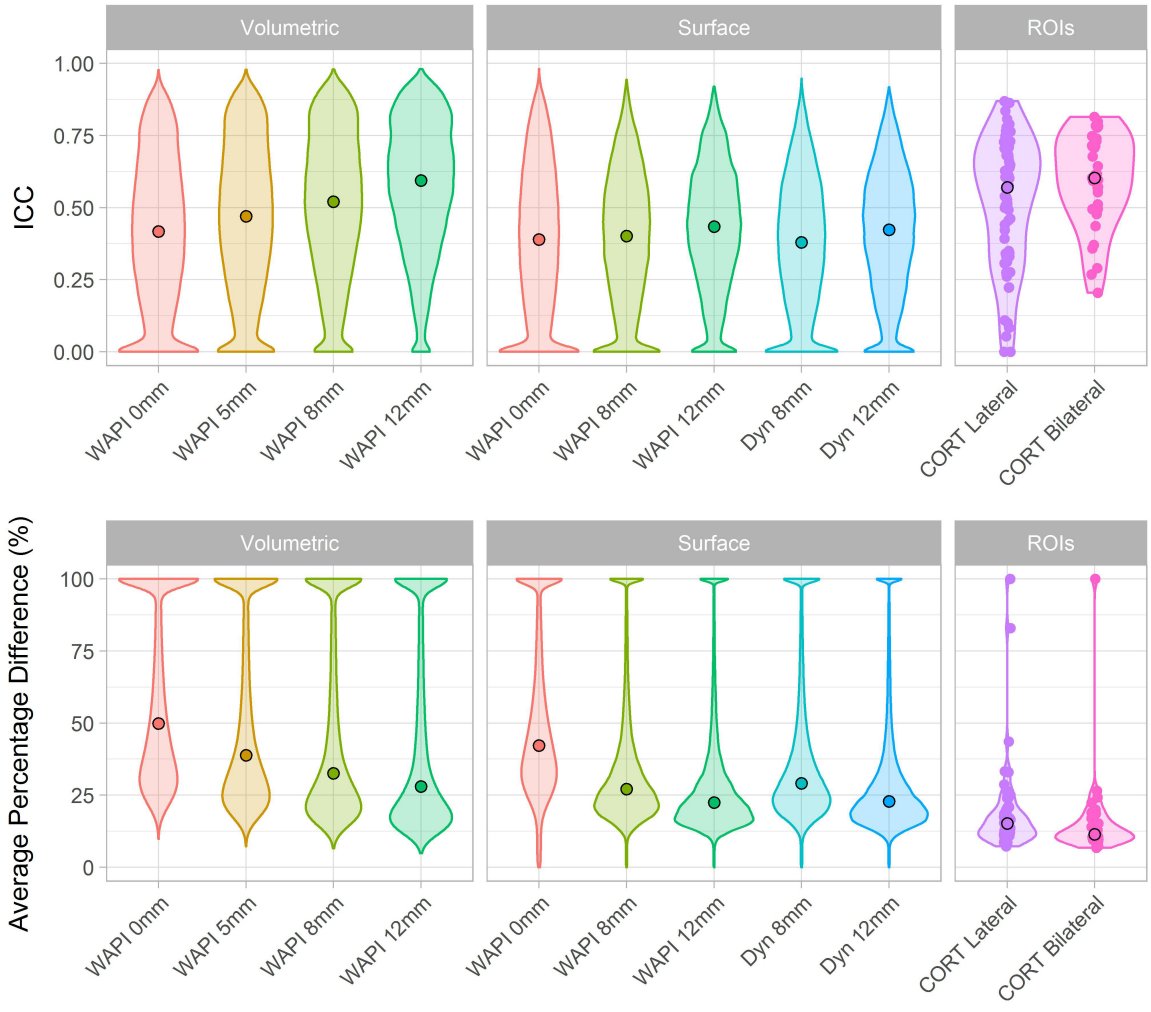

Figure 4.1: Comparison of voxel/vertex-wise test-retest value metrics for each method tested. Points in the centres of distributions represent medians. Values beyond the y axis limits have been truncated to be equal to the axis limit value in order to visualise the entire distribution of values.

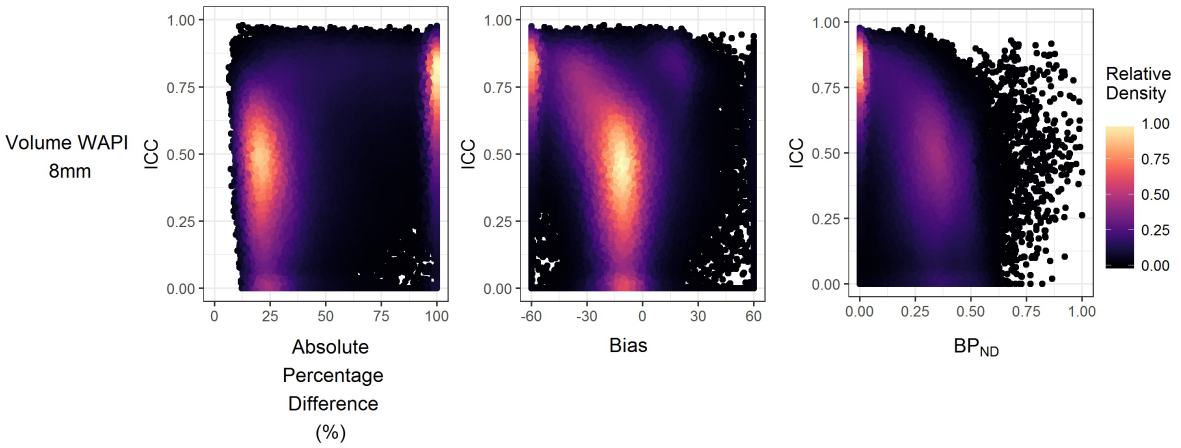

Figure 4.2: Comparison of voxel/vertex-wise ICC values and APD, Bias and BP_{ND}. Colours represent relative two-dimensional density estimates. Values beyond the axis limits have been truncated to be equal to the axis limit value in order to visualise the entire distribution of values.

A secondary conclusion from this study arose as a result of our having provided ROIs as a baseline comparison group, since ROI analysis is more common in PET imaging than parametric approaches. As expected, ROI analysis exhibited higher reliability and better repeatability, but the degree to which ROI analysis showed improvements over parametric analysis was unexpectedly large. We therefore recommend the use of surface-based methods for performing exploratory statistical parametric analysis of cortical regions, but only when the use of ROIs is not appropriate (i.e. when the expected effects may not be restricted to the anatomical boundaries of a ROI, and when no hypothesis can be made about their location).

Themes

Reliability

We used the ICC as a measure of reliability, as well as the related repeatability, to assess the performance of each method. Reliability was therefore the primary outcome measure for comparison of the performance of the different techniques.

Replicability

In this study, we performed a replication of the results of Greve et al. (2014)[40], using a different PET camera, a different PET tracer, a different kinetic model, as well as a different protocol

for the volumetric methods (which we believed might improve the performance of volumetric methods relative to surface-based methods). We successfully replicated the previous findings, despite the differences between the studies, suggesting that these previous findings are not only replicable, but also generalisable (as described in Figure 1.9)[100].

Reproducibility

This study constituted a valuable learning experience in reproducibility, as the processing pipeline required switching several times between the use of R and MATLAB, but these two were never integrated. However, after reviewers requested that we analyse the whole cortex as opposed to just the frontal cortex as we had originally, this could be achieved in a matter of days, as all steps following definition of the ROIs could be run from their scripts, and all graphs and figures were automatically created and added to the manuscript. This study was reproducible in parts, and this resulted in a great deal of time saved. The data and code were not, however, shared.

4.2 Study II: Reliability of [^{11}C]SCH23390 binding using different image processing methods

There exist an very large number of different methods for image processing of neuroimaging data which both can be, and are, applied, which limits the generalisability of findings[102,104]. In this study, we aimed to study whether the use of automated and manual ROI delineation methods, as well as frame-by-frame realignment for movement correction, altered the test-retest repeatability and reliability of BP_{ND} measured using [^{11}C]SCH23390. Automated ROI delineation methods have the advantage of being less time-consuming to apply, and theoretically unbiased, while manual ROI delineation should theoretically be more accurate, but has the possibility of being biased. Frame-by-frame realignment has the advantage of reducing the influence of small movements during the PET measurement, however it does involve reslicing each frame of the PET measurement, which induces a small degree of smoothing. For this reason, with young, healthy individuals who do not move greatly, it is possible that the process of performing the realignment of frames could introduce more noise than the small movements themselves do.

We found that frame-by-frame realignment resulted in higher BP_{ND} values as well as improvements in repeatability and reliability. Delineation of target ROIs using automated delineation methods resulted in lower BP_{ND} due to the ROIs being larger in size, however automated delineation of target regions also resulted in slightly higher reliability in all

regions, as well as lower COV. Furthermore, ROI sizes showed lower COV, indicating greater consistency. Automated delineation of the reference region resulted in improvements in repeatability and reliability, as well as slightly larger ROI volumes. Importantly, the delineation of the reference region was performed not according to anatomical guidelines, but rather with the best properties for quantification (see *Definition of regions of interest* within the *Materials and Methods* chapter). All comparisons of reliability showed inter-regional consistency of effects, but none were significantly different due to the large confidence intervals around ICC estimates.

Similarity in ROI delineations was examined using the Jaccard index, defined as follows:

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

In other words, the Jaccard index is equal to the volume where A and B intersect (i.e. the number of voxels which belong to both A *and* B), divided by their union (i.e. the number of voxels which belong to either A *or* B *or* both). Jaccard indices were high in the striatum, however they were low in the cortical regions as well as in the cerebellum. The differences within the cortex are explained by the fact that there is less agreement between exact definitions of what constitutes the dorsolateral prefrontal cortex for example. The differences in the cerebellum came about as a result of the automated method selecting more voxels within the lateral extent of the cerebellum, and the manual delineation coming close to the occipital cortex.

In conclusion, frame-by-frame realignment is important even in healthy control subjects, and we recommend its routine application to all measurements in the absence of movement correction prior to reconstruction. We further show that automated delineation shows greater consistency of its delineation, does not decrease the reliability or repeatability, and may indeed even improve them, and can be reasonably be relied upon to be less biased compared to manual ROI delineation. In combination with the fact that automated delineation requires substantially less researcher time, this method can be routinely applied instead of manual delineation. However, for studies in which there are expected to be gross morphological changes, the accuracy of automated delineation compared to manual delineation should be evaluated prior to its application. Choice of delineation of the reference region improved test-retest performance, presumably because of its caution by design: that it avoided regions of the brain which might affect reliable quantification by wide margins. This represents a promising approach for delineation of reference regions.

Themes

Reliability

We used reliability and repeatability to assess the performance of each method.

Replicability

In this study, we examined the test-retest reliability of [^{11}C]SCH23390, which has been examined before[237,238] in other studies, with slightly different image processing procedures, as well as different PET systems. We were able to obtain outcomes in a similar range to those of Hirvonen et al. (2001)[237]. Kaller et al. (2017)[238] observed much higher reliability in all regions, however the reason for this discrepancy appears to be the wide age range of individuals included in this study. This increases the between-subject variation, and thereby inflates the ICC values. Age can be incorporated as a regressor into ICC calculations, and in the case of having a standard deviation of age of over 10 years, probably should have been. In conclusion, however, in this study, I would consider our result to successfully replicate previous observations.

Reproducibility

This study, and the reprocessing of this data, resulted in the creation of tools for reproducible extraction and test-retest analysis which are currently publicly available in the *KI PET Tools* (https://github.com/mathesong/kipettools) and the *relfeas* (https://github.com/mathesong/relfeas) R packages respectively. This meant that all processing of data after image analysis could be performed reproducibly through subsequent iterations, although the raw data was not shared.

4.3 Study III: Reliability and validity of simplified ratio-based methods of quantification for [^{11}C]PBR28

Due to the fact that there is no reference region of the brain which is devoid of the translocator protein (TSPO), kinetic modelling using the metabolite-corrected plasma as the input function is considered to be the gold standard for [^{11}C]PBR28. V_T is the most commonly reported outcome measure from this approach, however this outcome exhibits a large degree of inter- as

well as intra-subject variability. This limits its sensitivity for the detection of effects in applied studies. Furthermore, collection of arterial plasma data is expensive, difficult or can even not be possible in some centres, or for some patient groups. For this reason, there is a need for methods of quantification of [^{11}C]PBR28 which do not require arterial sampling. These two considerations have led some groups to propose, as well as to use, ratio-based methods for quantification, i.e. the distribution volume ratio (DVR, $\frac{V_T^{target}}{V_T^{ref}}$) or the standardised uptake value ratio (SUVR, $\frac{\int SUV^{target}}{\int SUV^{ref}}$)[169,239,240]. These methods result in large reductions in both inter- as well as intra-subject variability. In this study, we aimed to assess the extent to which these ratio-based methods were reliable and associated with V_T.

We focused on the frontal cortex as a representative target region (and exhibited correlations greater than r=0.9 with all other regions for both V_T and SUV). We showed high reliability for V_T and moderate to high reliability for SUVs in both high- and medium-affinity binders (HABs and MABs). For ratio-based methods, we used two denominator regions: the whole brain and the cerebellum. The reliability of SUVR was found to be moderate (≈ 0.75 on average), while the reliability of DVR was poor (≈ 0.5 on average). SUVs were found to be fairly highly correlated with V_T (HAB R^2=0.64; MAB R^2=0.86). However, DVRs and SUVRs showed little to no assoication with V_T (all $R^2 \leq 0.34$ divided by genotype), with half of the associations showing negative correlation coefficients.

We examined this association further by comparing V_T and SUV values between ROIs to examine the correlational structure of the data. All regions were highly correlated (all V_T R>0.95, all SUV R>0.92). This suggests that ROIs are highly associated with one another on a pairwise level, but does not answer whether there is extra remaining information in the data after taking the main component of variation out. We therefore performed principal component analysis (PCA), a multivariate dimension reduction technique, to assess how much of the variance was remaining in the data after removing the first major dimension of variance. Using all six ROIs, the first component of the PCAs explained 98.7 and 99.4% of the total variability for PET1 and PET2 respectively. This suggests that almost all variation in all of the ROIs is explained by a single dimension of variance, and that almost nothing is left afterwards. Furthermore, from the test-retest analysis, this amount of variance is within the margin of test-retest error. This suggests that little to no biologically relevant signal likely remains after dividing the signal originating in one region with another.

We conclude that even if V_T is not a perfect outcome measure, if it is even at least moderately associated with TSPO levels in the brains of healthy control subjects, then the validity of ratio-based methods must be called into question. The poor reliability of the DVR provides further evidence that most of this signal is attributable to noise. In fact, even if the DVR were

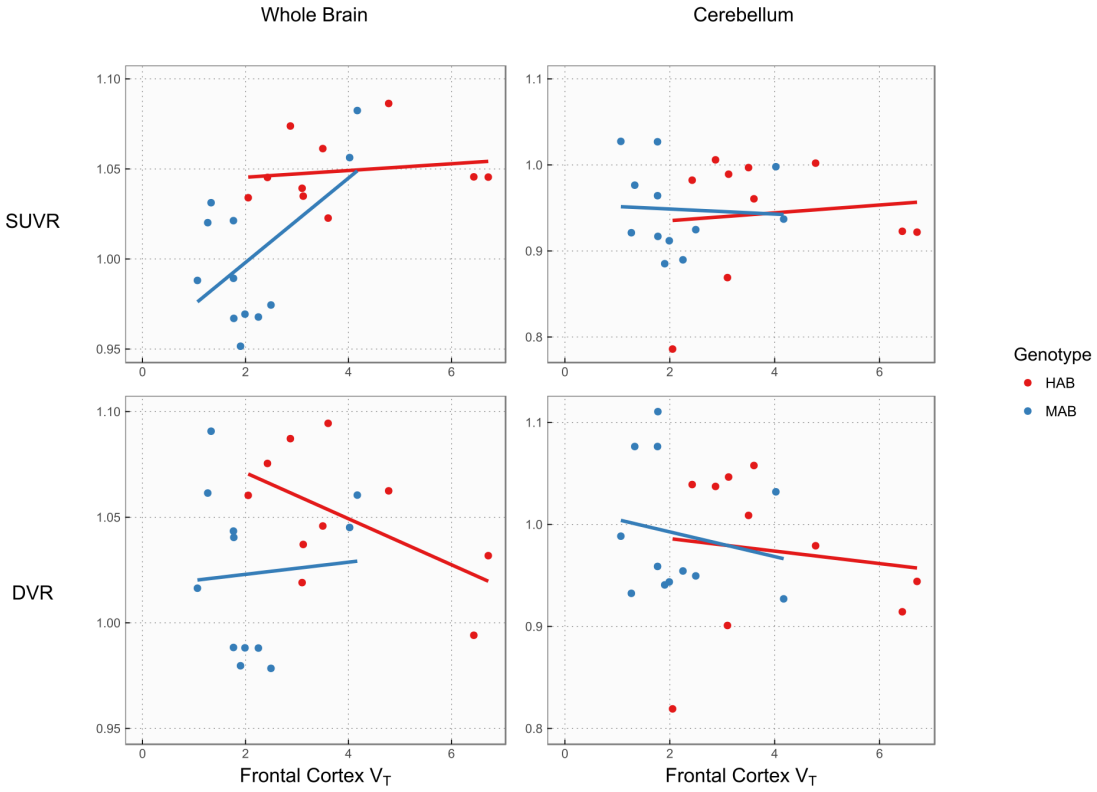

Figure 4.3: Associations between frontal cortex V_T and ratio-based outcomes, using the whole brain and cerebellum as denominator regions. Dotted lines indicate repeated measurements.

to be shown to be a valid measure of TSPO binding, the poor reliability of this measure implies that its utility would be questionable for comparing individuals. Lastly, we show why its reliability is so poor: almost all signal across all six ROIs, including the denominator regions, is attributable to a single dimension of variance with almost no residual signal remaining which might be explained by anything else. This provides a reasonable explanantion for why the outcomes of ratio-based quantification methods can primarily be attributable to noise. For data for which blood data is not available, we recommend that the SUV (and not the SUVR) might constitute a reasonably valid compromise outcome measure due to its high association with V_T as well as its reliability. However, the use of SUVs relies on the assumption of no differences in radioligand delivery between groups, which cannot be safely assumed in patient samples.

This study was limited by the fact that it was conducted in a sample of young, healthy individuals. For this reason, no regionally-specific alterations in TSPO availability are expected. The use of SUVRs and DVRs can certainly describe large, localised effects such as the incidental

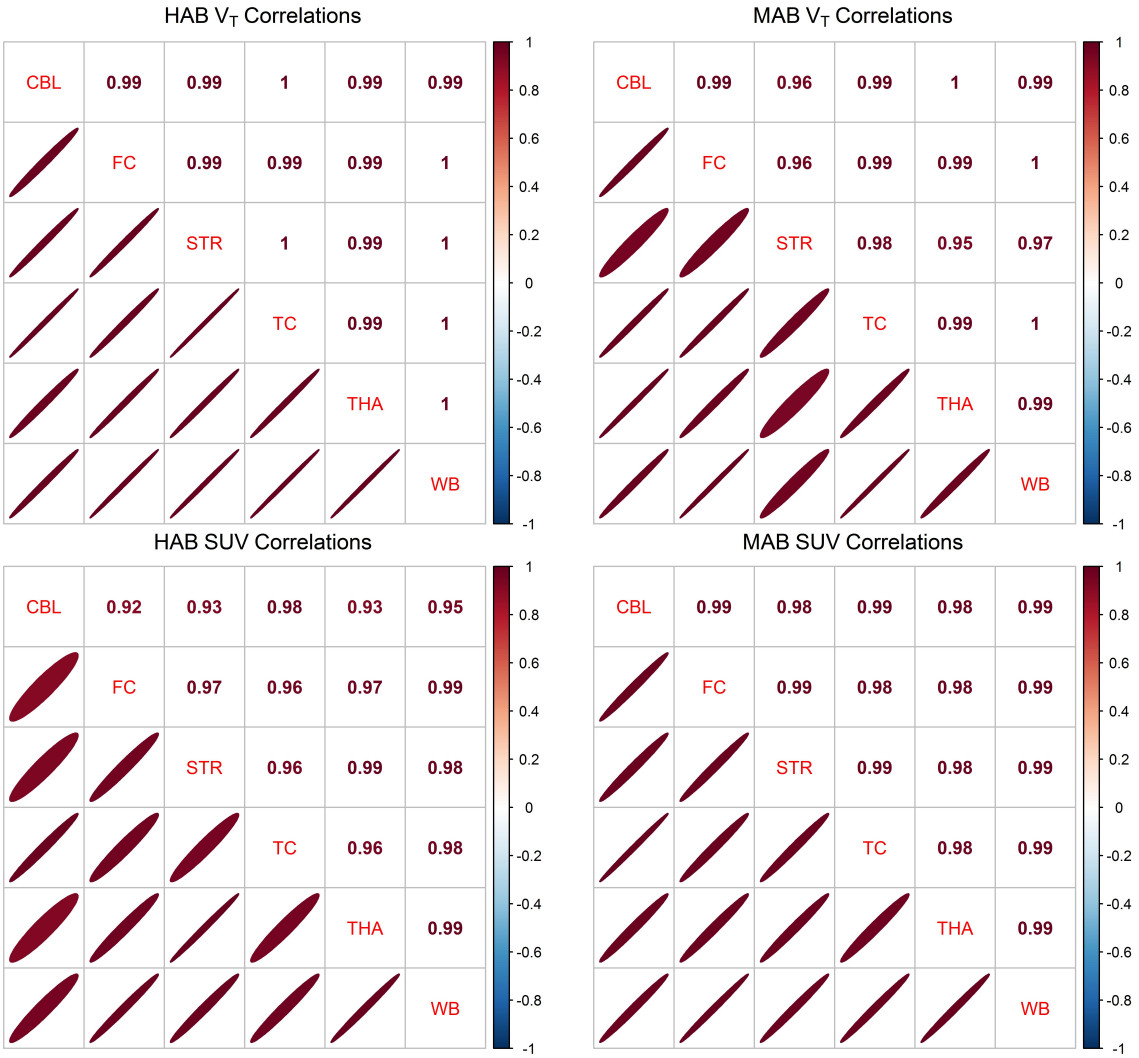

Figure 4.4: Interregional correlations of PBR28 V_T and SUV. Values represent Pearson's correlation coefficients. Ellipses designate the magnitude and the direction of the correlation values.

finding observed in Kreisl et al. (2009)[241]. For other conditions for which the effects cannot be assumed to be so large, or be assumed to be so local, significant equivalence should be demonstrated for the denominator regions which has not yet, to our knowledge, been shown in any study utilising these measures. Another concern is that of the reliability of the DVR: even if equivalence can be shown for the reference region, the results of this analysis suggest that the reliability of the DVR is insufficient for its application to answer clinical questions with small effects. However, the utility of DVRs increases with increased effect sizes. This was one of the primary motivations which led to Study V, where this question will be covered in greater detail.

Themes

Reliability

We used reliability and repeatability to assess the performance of each outcome measure.

Replicability

In this study, we examined the test-retest reliability of [^{11}C]PBR28 SUVs and SUVRs, which have been previously examined[240]. We obtained generally similar reliability despite our using a different, but more appropriate, variant of the ICC[94]. We also obtained generally similar repeatability (measured using the absolute percentage difference) to those presented in the erratum to this paper[242]. However, while we concluded that SUVR was likely not a particularly useful outcome measure, Nair et al. (2016)[240] performed power calculations suggesting that as few as 3 individuals would be needed in a longitudinal (i.e. pre-post) study to detect a 5% change in SUVR, with 90% power. While this was reported as being indicative of SUVR being a useful measure, what was not reported was that this 5% change represents a Cohen's D effect size of over 6 due to the extremely low degree of inter-subject variance after division by the denominator region. For comparison, an effect size of 0.8 is estimated for the increases in presynaptic dopamine function in schizophrenia which have been replicated many times[21]. We therefore do not consider an effect of this size reasonable, and do not consider this power analysis to be indicative of great utility for SUVRs.

Although not the main outcome of this study, another test-retest study of [^{11}C]PBR28 has also been reported[157]. While the outcomes of the present study were very similar to Collste et al. (2016)[158], which is to be expected due to both studies using the same data, Park et al. (2015)[157] observed substantially better repeatability (i.e. lower APD, by approximately 50%).

One potential reason for this is that their mean V_T values are approximately 30% higher on average, likely due to their use of motion correction prior to reconstruction.

Reproducibility

This study was completed in a completely reproducible fashion, with all figures and tables generated programmatically from the original time activity curve data. Linked and executable code and data in the form of a reproducible analysis report are shared online so that others may download the materials, run it for themselves to assess the robustness of our conclusions (https://github.com/mathesong/PBR28_RatioMethods).

4.4 Study IV: Diurnal and seasonal variation of the brain serotonin system

Circadian and seasonal changes in physiology are vital for the survival of animals and plants as they allow for physiological anticipation to the regular 24 hour day-night cycles of the earth. These changes take place in every cell, and are orchestrated in mammals by the master pacemaker, the suprachiasmatic nucleus (SCN). These rhythms are entrained to the environment primarily through natural rhythms of light exposure[243]. Circannual (seasonal) changes in physiology appear to have a circadian basis as a response to changes in day length[244]. Disturbances in chronobiology are thought to be involved not only in seasonal affective disorder[245,246], but also in non-seasonal major depressive depression[247,248]. Furthermore, both of these disorders have been shown to respond to chronotherapeutic treatments[249–251]. The serotonin (5-HT) system has attracted significant interest both in circadian regulation, but also in understanding mood disorders, and 5-HT concentrations have been shown to exhibit circadian and seasonal variation in both animals[252,253] and humans[254,255]. Even in plants, serotonin and melatonin exhibit both diurnal and seasonal rhythms as a function of the availability of solar energy[256]. Circadian and seasonal changes have also been observed in 5-HT_{1A} receptor (5-HT_{1A}R) and transporter (5-HTT) concentrations in animals[257,258]. Using PET imaging, seasonal changes in 5-HT_{1A}R[259] and 5-HTT[260–262] have been shown in humans, but circadian changes have not been reported. In this study, we aimed to attempt to replicate previous findings of seasonal changes in these proteins, as well as extend these findings to examine diurnal changes.

We defined a multiple regression model incorporating age, season (i.e. daylength) and day course (i.e. the fraction of the photoperiod which had elapsed at the time of PET) and applied

this to data from three regions for each radioligand. For each radioligand, the ROIs included the dorsal raphe nucleus (DRN) as well as two other sets of regions which were based on results from previous papers ([^{11}C]WAY100635: cortical and subcortical, and [^{11}C]MADAM: striatal and extrastriatal). For [^{11}C]WAY100635, we successfully replicated previous findings of increases in BP_{ND} in the summer months for both cortical and subcortical regions. Furthermore, we observed a positive correlation between BP_{ND} and elapsed day course. For [^{11}C]MADAM, we could not replicate previous findings of lower binding in the summer months[260,261] in any region, however we did observe a significant decrease with day course in the DRN.

This study was limited by the fact that it was cross-sectional, and based on old data, and we thereby did not have control over the distribution of times of measurement. The ideal experiment would involve performing PET imaging at various times across the day and night within the same individuals. However, this would also presumably require large samples as these effects do not appear to be large.

In conclusion, it is important to consider the potential influence of physiological processes such as circadian and circannual alterations on underlying biochemistry. These changes are important both from the perspective of examining differences between individuals in the 5-HT system as an experimental factor which should be considered, as well as for understanding the biological role of 5-HT itself in both healthy ciradian and seasonal physiology, as well in depressive disorders[245–248,262].

Themes

Reliability

As discussed more in Study V, one of the assumptions made in assessing the reliability of an outcome is that the underlying levels will remain the same from test to retest. Diurnal variability in protein availability measured using PET has been observed for metabotropic glutamate receptor subtype 5[263], and preliminary diurnal fluctuations have been reported for TSPO[158]. This poses an interesting edge case for the assessment of reliability in PET studies, for which another ICC formula might be appropriate which additionally takes systematic differences between the two measurements into account[94].

Replicability

In this study, we successfully replicated the results of Spindelegger et al. (2011)[259], however we were not able to replicate the seasonal changes in 5-HTT which had previously been shown[260,261]. McMahon et al. (2016)[262] suggested that this may have been a result of our not having included the polymorphism in the serotonin-transporter-linked polymorphic region (5-HTTLPR) in the model. We also suggested that this may be a result of low power, as we only had access to a sample of 40 individuals for the [^{11}C]MADAM analysis. It should be noted that another longitudinal study examining 17 seasonal affective disorder patients and 23 healthy controls with low seasonality in both the summer and winter, showed that patients had higher 5-HTT binding in the winter months compared to controls[262]. However this study also observed *lower* 5-HTT binding in the healthy control group in the winter months compared to the summer months, which is the opposite direction to previous findings. Future studies will be needed to shed light on the nature, size and direction of any such changes.

Transparency Statement

This study was conducted at an early stage of the PhD learning process. There are therefore several parts of the analysis which could have been conducted in a better manner now.

The plots in this paper make evident a significant issue of heteroskedasticity in the assessment of the effects of day course on [^{11}C]WAY100635 binding. The two individuals who appear to be influential outliers were in fact twins, suggesting that their low values are likely to be due to some other factor.

This analysis is only one from the garden of forking paths, and the other was not reported. We had initially included all ROIs separately, and had reported assocations of [^{11}C]WAY100635 BP_{ND} with the number of hours elapsed since sunrise (i.e. not the fraction), and of [^{11}C]MADAM with elapsed day course fraction. Based on reviewer comments about a problem of multiple comparisons, we combined ROIs into larger regions. Because we believed that the elapsed day course fraction was the more valid metric, we tested this first, and observed significant associations for both tracers, and we therefore did not proceed to test the hours elapsed since sunrise. However, as Gelman & Loken (2013)[264] argues, the garden of forking paths can be a problem even when there is no p-value fishing expedition, since if there were to have been no result, then the expedition might have continued. It was likely only because the value was significant that we did not continue to probe the data.

Three [^{11}C]WAY100635 and one [^{11}C]MADAM measurement were excluded from the analysis.

Two of the [^{11}C]WAY100635 measurements were excluded *a priori*, however the remaining two measurements were excluded based on image artefacts which were only discovered due to the values for the DRN lying far from the regression line. This does not mean that the measurements were correct - they did contain image artefacts after all - but our having given greater scrutiny to those measurements which did not conform to our hypothesis can be considered to be a QRP.

A further issue is that of partial HARKing (Figure 1.7): in this study, we initially set out to examine seasonality and the relationship of 5-HT proteins with measures of light assessed using data from the Swedish Meteorological and Hydrological Institute. Examining circadian changes was initially a secondary aim. Over the course of the investigation, this aspect of the investigation grew more important, presumably as a result of its being more compelling based on the data, but also perhaps as I learnt more about the biological basis of seasonality. To separate these two after the fact is difficult due to hindsight bias.

4.5 Study V: We need to talk about reliability

Measuring and understanding the reliability of outcome measures is extremely important, as it allows us to gauge whether the accuracy of our data is capable of meaningfully distinguishing between individuals. Not accounting for the reliability of outcome measures increases the risk of type II errors, or can be considered to increase the risk of type M (magnitude) and type S (sign) errors[265]. Reliability of outcome measures is associated with degree of attenuation of the expected effect size. For a technique such as PET, which is so expensive, it is vital to have an estimate of the expected effect size in order to make an assessment of the feasibility of a study to answer the specified research question with the given resources: failure to account for these factors can result in large wasted costs as well as the exposure of research participants to radioactivity unnecessarily. For PET imaging, reliability is typically assessed using test-retest studies with young, healthy control subjects. However, as described before, reliability is sensitive to the amount of variation in a sample, and this must be taken into consideration when assessing whether a specified outcome measure is likely to be useful for comparing individuals in a new sample. For example, this issue limited our interpretation of the results of Study III, where we were unable to exclude the possibility that ratio-based outcome measures which showed poor reliability in our sample of young healthy controls, may indeed show good reliability in certain clinical samples. In this study, I describe a method for how to extrapolate an approximation of reliability for new, different samples based on the summary statistics of published papers, as well as step through five case studies to demonstrate

how these calculations can be performed, as well as how reliability can, and should, be used in the planning of new studies.

Because reliability is a function of the relative variance attributable to measurement error and true inter-individual variance (equation 1.8), the reliability can improved either by reducing the measurement error, or by increasing the amount of true inter-individual variation in the sample. As a result, using an ICC value calculated for a small, uniform, test-retest sample, a new ICC can be estimated for a new sample of individuals whose inter-individual variation is higher, by substituting the variation of the new sample into the equation, and assuming that the measurement error is constant between the studies. In fact, even if the measurement error is expected to increase or decrease in the new sample, the extent to which this is expected to occur can be approximated and incorporated into the expression (equation 4.1).

$$ICC_{NewStudy} = 1 - \frac{(\rho \times \sigma_{TRT})^2}{\sigma^2_{NewStudy}} = 1 + \frac{\rho^2 \sigma^2_{TRT}(1 - ICC_{TRT})}{\sigma^2_{NewStudy}} \tag{4.1}$$

where ρ represents the error inflation factor, which is the multiplicative increase in expected measurement error in the new study (i.e. $\rho = 1$ assumes equal error between studies). Using this equation, we can thereby calculate the required standard deviation in a new sample which would be required to reach a suitable level of reliability. If we know the required standard deviation for this to be the case, we can calculated the size of the difference between the groups which would be required between two groups (in the case of a t-test) for such a standard deviation to be reached.

The relationship of effect sizes with the reliability of measurement can be described by the equation 4.2 (which is often called Nunnally's equation[266], although it has a much longer history)[267,268].

$$r_{ObsA,ObsB} = r_{A,B} \times \sqrt{reliability_A \times reliability_B} \tag{4.2}$$

This is demonstrated in Figure 4.5, which shows the attenuation of underlying effect sizes as a result of stochastic variability in the measured outcome described by the reliability. In the figure, I refer to the population effect size, which is the true effect size of the distributions from which the individuals are sampled from, and the sample effect size, which is the effect size obtained based on the sample obtained from the population.

Based on the instability of effect size estimates, as well as the degree to which individual measurements are not representative of underlying reality with low reliability values, Nunnally

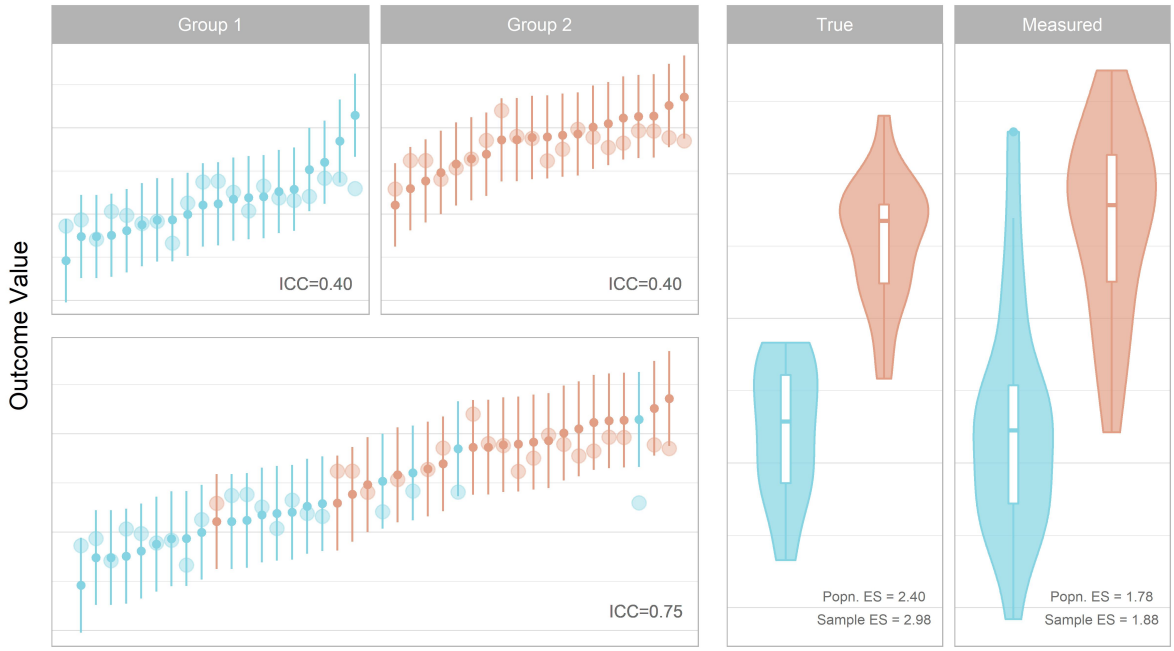

Figure 4.5: Left: Measured values and their 95% confidence intervals are represented by the small points and error bars. Underlying true values are represented by the larger points. Low reliability within groups is increased for the total sample in a comparison, given a sufficiently large effect size, due to the larger variance of the combined sample. Right: True underlying effect sizes (Cohen's D, left) are attenuated by measurement error (right). The population (Popn.) effect size (ES) and sample ES are of the underlying distributions from which the data are sampled and of the obtained sample respectively.

(1987)[95] recommended a reliability of 0.7 as a default lowest acceptable standard of reliability for scales used in basic research, and that 0.8 should be seen as adequate. For applied settings in which important clinical decisions are made based on measured individual outcomes, he suggests a reliability of 0.9 as a minimum and 0.95 as adequate, since even with a reliability of 0.9, the standard error is almost a third the size of the standard deviation. These recommendations were made referring to internal consistency in psychometric measurements though, and not test-retest reliability. Traditional standards for test-retest reliability, on the other hand, have been more liberal, with Fleiss (1986)[96] suggesting that ICC values between 0.4 and 0.75 could be considered good. These recommendations were made based on psychometric questionnaires, for which individuals can be expected to change their opinions over time. For PET imaging, this source of variation can, in most cases (although Study IV is a good example where this might not strictly be true), be considered to be negligible. For this reason, I argue that the recommendations of Portney & Watkins (2015)[98] are more applicable to PET research due to their closer correspondence with those of Nunnally et al. (1978)[95] defined for internal

consistency.

I will briefly cover the results of the case study examples. In *example 1*, I demonstrate how a measure which showed unacceptable (0.32) test-retest reliability can be successfully applied in a sample with higher inter-individual variation, for which it exhibits an estimated reliability of 0.93. In *example 2*, I demonstrate how accounting for the (acceptable) reliability of two measures during power analysis can nearly double the number of participants needed in a study to test a specified relationship. This can be expected to render many studies incapable of answering their research question with the available resources. Furthermore, this can help researchers to avoid wasting resources, as these conclusions can be reached before even beginning to conduct the study. In *example 3*, I show how large, robust within-individual effects do not necessarily allow for the comparisons between individuals using these measures, and how this can be approximated using measures of reliability and standard deviations from summary statistics. In *example 4*, I show that high reliability does not imply sensitivity for small proportional effects, and how it is helpful to consider effect sizes rather than percentage changes. Finally, in *example 5*, I return to the question from Study III regarding how large the differences would be required to be to render the DVR an effective comparison measure. I show that, according to our data, for the DVR to reach the lowest acceptable reliability for basic research according to Nunnally (1978)[95] (i.e. 0.7), effect sizes would be required to be between very large and huge[269], and to reach acceptable reliability, would be required to be 25% larger than the huge guideline. For such large effects, the more easily interpretable, and much less controversial, outcome measure of V_T will be more than capable of distinguishing between groups, thereby rendering the DVR practically obsolete for between-groups comparisons.

An important divide between reliability and repeatability should be emphasised here: reliability is a measure of the ability of an outcome measure to distinguish *between individuals*, while repeatability is a measure of the ability of the consistency of an outcome measure across repeated examinations *within the same individual*. Hedge et al. (2017)[270] recently showed that some of the most well-validated cognitive tasks, such as the Stroop or Go-NoGo tasks, despite exhibiting extremely robust within-individual effects, showed extremely poor reliability for inter-individual comparisons, primarily attributable to low variation. They note that these characteristics can be traced to a common underlying characteristic of low inter-individual variation. Historically there has been a separation of within- and between-individual research, both of which have prioritised different aspects of the distribution of the outcome measures. They recommend caution in the assumption that robust outcomes from one paradigm should apply to the other, since, as a heuristic, the opposite is more often likely to be true (although this can be the case with extreme differences both between and within individuals).

In conclusion, this paper motivates the use of reliability for better prospectively gauging the feasibility of performing new between-individual studies with the resources available. The paper also describes a method for estimating reliability for new, different samples, based on the reported summary statistics from previously conducted test-retest studies.

Themes

Reliability

This paper is centred around reliability, application of reliability for power analysis, as well as estimation of reliability for new samples.

Reproducibility

This paper was written in an entirely reproducible manner, where any change in parameters or in functions would be directly incorporated into the figures and into the text. The *relfeas* R package which can be used to implement all of the functions described is available online (https://github.com/mathesong/relfeas), and is accompanied by code to implement all of the calculations made in all of the case studies.

4.6 Study VI: Serotonin 5-HT_{1A} receptor binding and self-transcendence

As discussed in the previous study, low between- and within-individual variability tend to be beneficial for within- and between-individual study designs respectively[270]. The high degree of inter-individual variability in neuroreceptor densities between healthy controls, both *post mortem*, and later in vivo using PET, was initally seen as problematic for comparisons of patient and control groups, since this diminished the power of PET for examining small proportional changes in neuroreceptor concentrations between groups[8]. However, this also presented an opportunity to examine the sources of this inter-individual variation. The study of personality traits is, by definition, related to inter-individual differences in stable patterns of behaviour, cognition and emotion. Coupled with the fact that these traits have shown a high degree of heritability[271], and are thought to be important predictors of psychiatric disorders at their extremes[272,273], the study of correlations between brain neuroreceptor densities and personality traits has flourished[8]. Sample sizes, however, have tended to be rather small, and

as such it is especially important that they are validated by replication. One highly influential study within the field of brain neuroreceptors and their association with personality traits is that of Borg et al. (2003)[274], which found strong negative associations between 5-HT_{1A}R BP_{ND} measured with [^{11}C]WAY100635 and the Self-Transcendence scale and its Spiritual Acceptance subscale from the Temperament and Character Inventory (TCI)[183]. This study was only conducted in fifteen men, however, and a previous replication attempt in a sample of 20 healthy controls (as well as 19 patients with major depressive disorder) did not find evidence of this relationship. This previous study, used frequentist statistical methods to analyse this relationship, which is not optimal for such an analysis due to the possibility of type M errors in the original study[265] which could lead to type II errors in the replication study.

We studied the relationship between 5-HT_{1A}R BP_{ND} measured with [^{11}C]WAY100635 using a larger sample (50 healthy men) of similar age to the original study, collected using the same PET camera as the original study, and using the same Swedish-language version of the TCI[223]. We used replication Bayes Factors (BFs) to assess replication success, as well as default BFs to assess the evidentiary weight of the new data as if the original study had never been published. We used these methods to analyse our data, as well as to re-analyse the results of Karlsson et al. (2011)[275]. Collectively, we showed moderate to strong evidence of a failed replication in both data sets, as well as moderate evidence for no association between the scales and [^{11}C]WAY100635 BP_{ND}. Even where evidence did not meet the level defined as 'moderate'[69,276], it still favoured the null hypothesis in all cases.

We also re-examined the original study in more detail, taking into account the realisation in recent years that statistical procedures which may previously have been considered acceptable, are able to generate false-positives at a rate higher than most appreciated[79]. Using the positive predictive value, a measure of the likelihood that a given research result is true given several assumptions[76,78], we arrived at a probability that the original finding was true of only 9%. This assumes a large effect size (i.e. r=0.5)[277], a pre-study odds of 10% which we believe to be reasonable for studies of neuroreceptors in personality, and a type I error rate of approximately 0.5 (due to the 21 comparisons made, which we considered to be equal to approximately 10 *independent* comparisons).

In conclusion, we could not successfully replicate the findings of Borg et al. (2003)[274], and showed consistent evidence for there being no association between [^{11}C]WAY100635 BP_{ND} and these scales. We thereby consider that we successfully replicated the results of Karlsson et al. (2011)[275]. While we cannot exclude the possibility that there is a true underlying effect and that both latter studies did not detect it, this result is considered to be unlikely based on the positive predictive value of this result. A more likely conclusion is either that the original

study made a type I error, or that the original study made a very large type M error.

Themes

Replicability

This paper represents a direct replication of a previous published, and highly influential result[99]. We show evidence for the original finding either being a false positive, or a substantial overestimation of the true underlying effect size.

Reproducibility

The code used to calculate replication BFs as well as to reproduce the plots is provided openly online (https://osf.io/x9gjj/).

4.7 Study VII: Translocator Protein in Patients With Psychosis: A Meta-analysis

The results of studies of TSPO in schizophrenia and psychosis have been mixed. These studies are limited, however, by their heterogeneity. Studies differed not only by the different tracers used and the medication status of patients, but also by outcome parameters, and accounting for genotype. As described earlier, since there is no region of the brain devoid of TSPO, kinetic modelling using the metabolite-corrected plasma as the input function is considered to be the gold standard. While V_T was reported in some studies[166–168,170], other reported outcome parameters have included V_S[160], BP_{ND} (calculated using rate constants)[161,165], pseudo-BP_{ND} (using reference-tissue models and a pseudo-reference tissue)[162–164], and the distribution volume ratio (DVR)[169]. TSPO genotype is also critically important, as it induces large differences between groups for second-generation tracers. The influence of genotype on the binding of first-generation tracers is controversial, since *in vitro* studies have suggested no differences between genotypes[278], while *in vivo* studies have shown genotype effects in peripheral organs of the body[279].

A further limitation of these studies is their size: all of these studies making use of arterial sampling, have been small (n<20 patients). None of these studies individually has more than

40% power to detect a medium-sized effect (i.e. Cohen's D=0.5). This severely limits the range of effects which could be detected.

Meta-analytic models allow researchers to compare the results of several studies and to derive a pooled estimate, which should hopefully be the best possible estimate of the underlying truth. The assumption of this model, then, is that the separate effect sizes are approximations of a single underlying 'true' effect size. However, when different studies have used outcome measures whose reliability differs drastically, then the effect sizes using the low reliability outcome measure can be expected to be underestimated[267]. The signal-to-noise ratio of the first-generation tracer [^{11}C]PK11195 is thought to be very low[152,280], and all of the outcome measures used in the above studies of schizophrenia patients have been shown to exhibit poor reliability[154]. For this reason, studies using [^{11}C]PK11195 should yield much lower estimates of the effect size compared to second-generation tracers. We therefore aimed to restrict our meta-analysis to studies employing second-generation TSPO radiotracers, making use of V_T as an outcome measure using metabolite-corrected arterial plasma as input function, and taking TSPO genotype into account. While meta-analyses are usually conducted using aggregate data (i.e. published effect sizes), we instead performed an individual participant data (IPD) meta-analysis. This means that data was obtained from previous studies at an individual participant level, which is considered to be the gold standard of evidence synthesis methods[281]. This further meant that the study emplying DVR[169] could be used, since we could request the V_T values which were used to calculate the DVR. This resulted in five studies which could be pooled. All authors of these studies agreed to be a part of the study, and sent raw V_T values for the frontal cortex, temporal cortex and hippocampus, as well as age, Positive and Negative Syndrome Scale (PANSS) scores, duration of illness and drug status (medicated, drug-free, drug-naive) at the time of measurement.

Before receiving the data, we pre-registered an analysis plan. This means that we decided on the precise statistical models which would be employed, how the model would be selected for performing inference, as well as how the comparison would be made. We also made this pre-registration plan publicly available in a version-controlled online repository, such that all modifications can be associated with their date of implementation. We performed standardisation (centring and scaling to have mean 0 and standard deviation 1) of V_T values within each genotype, within each study. This thereby adjusted V_T values from each radioligand and genotype into the same scale units (i.e. units of standard deviations, or in other words, into units of Cohen's D effect size) (Figure 4.6)). We performed the meta-analysis using Bayesian linear mixed effects modelling. We compared four models of increasing complexity using information criteria to select the model with the best fit to the data with which to perform inference. Inference was performed using Bayes Factors to compare the relative likelihood of

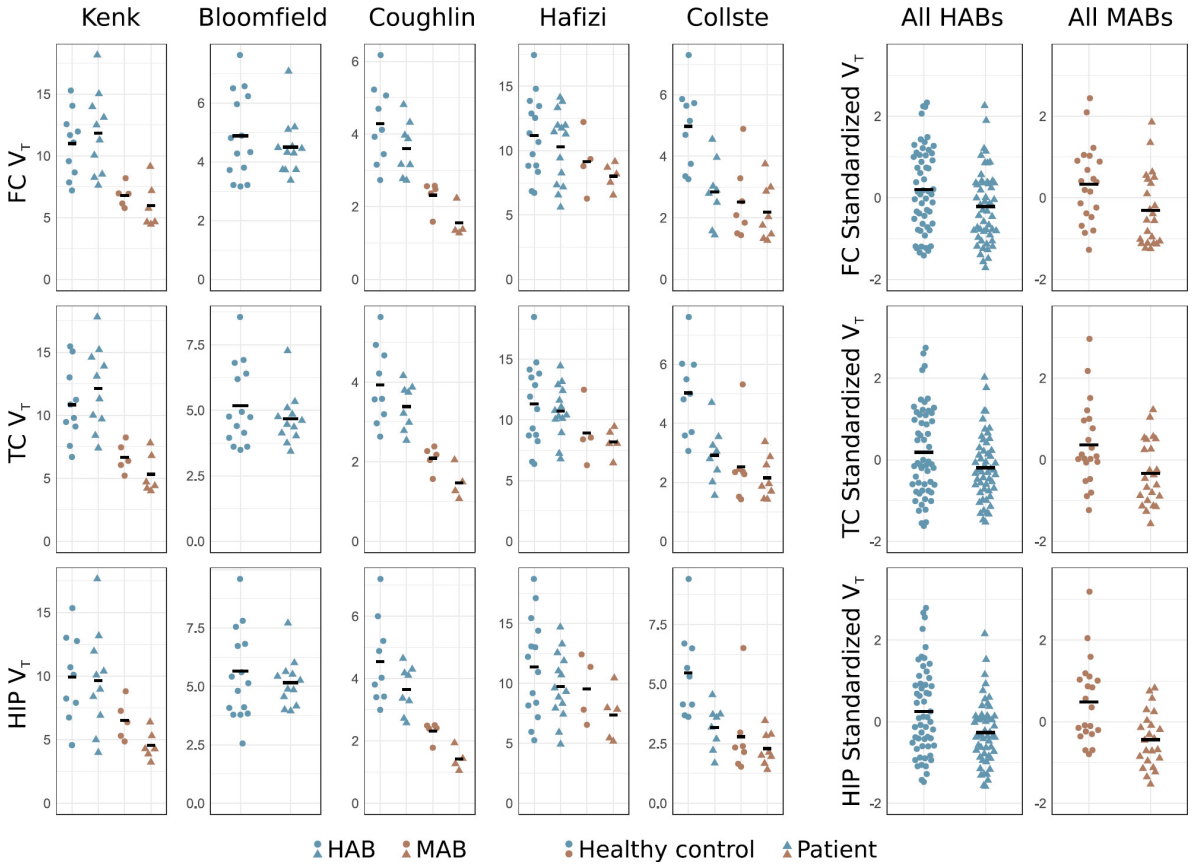

Figure 4.6: Left: V_T values from each of the constituent studies, shown separately for high- and medium-affinity binders, for the frontal cortex (FC), temporal cortex (TC) and hippocampus (HIP). Right: Standardised V_T values after pooling.

the data under each of three hypotheses: 1) no effect, 2) higher TSPO binding in schizophrenia, and 3) decreased TSPO binding in schizophrenia. We defined the hypothesis using half-normal priors over the difference between patients and controls with a medium expected effect size (Cohen's D=0.5) defined as the standard deviation.

We found that the simplest model was preferred, i.e. the model which assumed that all effect sizes from all studies and from all genotypes came from the same distribution. Performing Bayesian hypothesis testing using this model, we showed extremely strong evidence for decreases in TSPO levels compared to increases (decreases over 400 times more likely for all three regions), and strong evidence for decreases compared to the null hypothesis (decreases over 30 times more likely for all three regions). Performing subsequent parameter estimation, we showed that the posterior distribution was centred around a moderate effect size (≈ 0.5), although its credible intervals ranged between small and large. We also examined the effect of medication, the association between binding and symptom scores as well as duration of illness. The

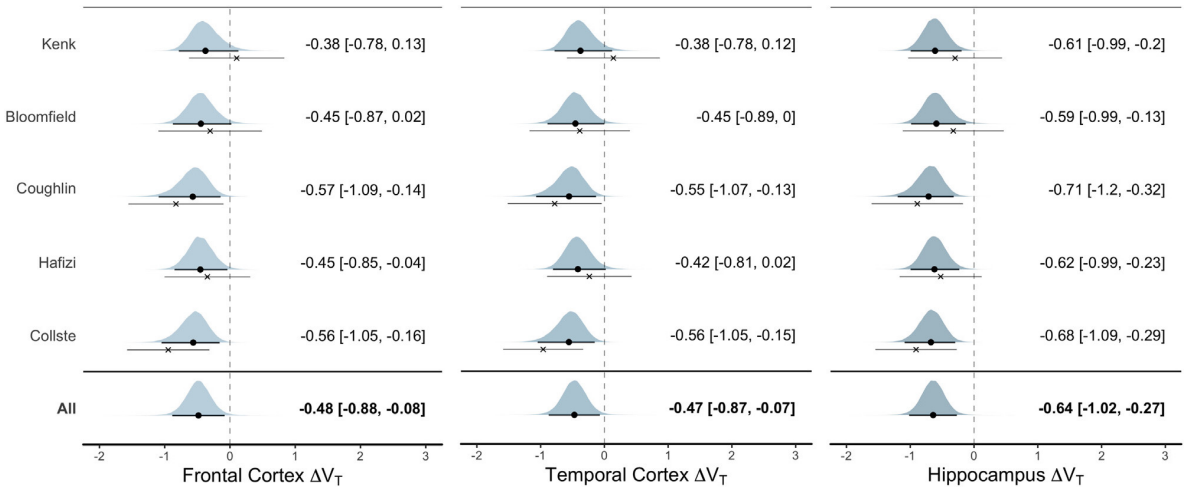

Figure 4.7: Standardised differences in TSPO V_T between patients with psychosis and healthy control subjects. For each study, the posterior probability density is shown in blue, the thick line depicts the 95% credible interval, with the black circle depicting the posterior mean (also written to the right of each plot). The cross and the thin line represent the raw patient-control mean difference and its 95% confidence interval for each study.

posterior probability densities showed that these factors had little to no effect on measured TSPO binding.

In summary, we showed strong evidence for decreases in TSPO binding in schizophrenia. This is contrary to the original hypothesis of higher TSPO binding in patients compared to controls. There has been extensive evidence for increases in pro-inflammatory markers in schizophrenia patients[142,282], and these results appear to contradict these findings. However, a recent translational study showed that increases in pro-inflammatory cytokines were associated with lower TSPO expression in the same regions of the brain[283]. Microglia and astrocytes, both of which contain TSPO, can exist in both pro- and anti-inflammatory states, and in vitro studies suggest that the pro-inflammatory macrophages and microglia may in fact be associated with lower TSPO expression in humans[284,285]. TSPO may, therefore, not be an exclusively pro-inflammatory marker in the human brain, and the lower levels reported may indicate a compensatory mechanism for a pro-inflammatory state, altered function of glial cells, or reduced density of immune and glial cells. In any case, regardless of the direction of change, TSPO binding appears to systematically differ between schizophrenia patients and controls, providing further evidence for the immune hypothesis of schizophrenia.

Themes

Replicability

This study set out to aggregate data over all of the previous studies corresponding to the inclusion criteria to arrive at the conclusion most supported by the sum total of the data. In so doing, we replicated previous results showing decreases[170], failed to replicate previous results showing increases[169], and observed an effect size which none of the previous studies were adequately powered to detect. This suggests that the studies which did not find significant differences can be considered to be indicative of insufficient evidence, rather than evidence of absence of an effect. Furthermore, we pre-registered our analysis plan in order to minimise the potential influence of biases, which have been associated with poor replicability[77,84,85]. A recent paper explains that meta-analyses are particularly susceptible to these biases, as subtle changes to inclusion criteria can have large consequences, and that for this reason, preregistration of meta-analytic protocols is arguably even more important than it is for clinical studies[286].

Reproducibility

Although it was not possible to share the V_T values which were used for the study, all code used to perform the analysis is uploaded to a public repository to ensure transparency. This includes the code used to query PubMed for the relevant articles and the specific search terms, to perform the analysis, as well as to generate all the figures in the paper is openly provided online (https://github.com/pontusps/TSPO_psychosis).

4.8 Study VIII: Delusional ideation and D1 receptor availability

PET studies comparing dopamine D1R levels between schizophrenia patients and controls have been inconclusive, showing higher levels[127,130,131], lower levels[126,129,132], and no difference[128] in levels of D1R in the frontal cortex. As discussed in the introduction, these studies have been limited by small sample sizes, and hence low statistical power, but interpretation of these studies have also been complicated by several important differences between studies. These include different tracers, different populations, duration of illness, but also, most importantly, medication status. In vitro studies suggest that exposure to antipsychotic medication may

result in decreases in D1R availability[133,134]. For this reason, it is important to study the differences in D1R availability at early stages of the illness, before antipsychotic treatment, or preceding its onset in order to understand its relationship to the disorder. In this study, we investigated the latter question.

We examined delusional ideation as a measure of psychosis proneness as described in the introduction. In order to make use of already collected PET and psychometric data, we developed a new scale for delusional ideation using items from the self-transcendence subscale of the TCI[183]. This scale was constructed using previously collected data from participants who did not undergo PET. In order to make use of already collected PET data, we contacted previous participants who had undergone PET with [^{11}C]SCH23390 in previous studies by letter and asked them to complete the 21 item version of the Peters Delusion Inventory (PDI) scale[171,172] online. And we also acquired new PET measurements with [^{11}C]SCH23390, and asked participants to complete both scales. As a result, this study made use of data collected from three separate cohorts of individuals, from four separate data collections. We divided this study into an exploratory, i.e. hypothesis-generating, component and a confirmatory, i.e. hypothesis-testing, component[84,85].

In the exploratory analysis, we successfully created a new scale for delusional ideation from items of the TCI self-transcendence subscale with adequate reliability (Cronbach's α=0.76). We called this scale the TCI-DI (delusional ideation) scale. We then assessed the correlation of this scale with PDI scores in individuals who completed both scales. The two scales were highly correlated (Pearson's R=0.64), thereby demonstrating convergent validity. Next, we aimed to assess the relationship between this scale and [^{11}C]SCH23390 BP_{ND} in the previously collected data. This data contained individuals with a bimodal age distribution, males and females, and there were originally two versions of the TCI-DI scale. Choosing one optimal analysis strategy was not possible, and would be essentially random (e.g. inclusion of old individuals versus inclusion of age in the regression model). For this reason, we used a multiverse analysis[224], by which all outcomes from all potential reasonable analysis decisions were presented in order to increase transparency as well as minimise the influence of the garden of forking paths[264]. From this analysis, we showed that all analytical approaches resulted in a negative association between TCI-DI scores and [^{11}C]SCH23390 BP_{ND} in both the DLPFC and the striatum.

In the confirmatory analyses, we first performed a replication study of our exploratory findings. For this, we used the replication BF (as employed in Study VI), with which we obtained moderate and strong evidence in favour of the null hypothesis in the DLPFC and striatum respectively. Next, we applied Bayesian hypothesis testing to compare the likelihood of hypotheses of a positive and negative association of [^{11}C]SCH23390 BP_{ND} with PDI scores

(i.e. increases or decreases in D1R availability with increasing delusional ideation). We made use of regularising, zero-centred prior for the main effect of interest, and defined informative priors derived from the literature to account for the effects of age on both [^{11}C]SCH23390 BP_{ND} and PDI scores. This analysis showed that the data were 5 to 10 times more likely to have occurred under the null hypothesis than either the increase or decrease hypotheses in the DLPFC. Similar results were observed in the striatum. In the final substudy, we fit the same model in a new sample, using the posterior probability destributions from the previous comparison as priors to obtain updated estimates of the size of the effect. This data brought parameter estimates even closer to zero for the association between [^{11}C]SCH23390 BP_{ND} and PDI scores. A 5 point difference in PDI scores has previously been found to be the difference between delusional patients and controls[172]. We showed that a change of this magnitude was associated with a change in BP_{ND} of only 1.5% of the mean, whose 95% credible interval did not exceed 8%. According to these estimates, a difference even of all 21 points of the PDI scale would only be associated with a difference in BP_{ND} of 6.5%.

In conclusion, despite the promising results obtained from the exploratory analysis, we conclude that there does not appear to be a linear association between D1R measured with [^{11}C]SCH23390 BP_{ND} and measures of delusional ideation. There are four primary potential explanations for these findings, namely that i) there is no association between D1R and psychosis, ii) the changes in D1R may only occur at the onset of the disorder, iii) that changes in D1R availability may be associated with some other aspect of psychosis proneness, such as genetic risk factors or other behavioural traits, or iv) that the association between D1R levels and psychosis proneness is of greater magnitude for higher levels of delusional ideation (i.e. the association is not linear).

Themes

Reliability

In this study, we created a new psychometric scale for assessment of delusional ideation from items belonging to the TCI scale, and selected items to obtain a good balance of reliability and face validity. We also used the PDI scale, which has previously been shown to be reliable[172]. In each separate investigation using either of these scales, we assessed the reliability of the measure in the specific sample, to make sure that it was capable of differentiating between individuals.

Replicability

In this study, we made use of both exploratory and confirmatory testing. Because exploratory investigations are prone to both false positives and overestimation of effect sizes, we made use of a confirmatory replication analysis to assess the veracity of these findings. The replication analysis, along with the remainder of the confirmatory analyses performed, suggests that the original exploratory findings were likely to have been a false positive.

Reproducibility

Although it was not possible to share the TACs or BP_{ND} values which were used for the study, all code used to perform the analysis is uploaded to a public repository to ensure transparency.

Transparency Statement

It is noteworthy that the exploratory finding of a negative association between TCI-DI scores and [^{11}C]SCH23390 BP_{ND}, despite our transparently reporting the full extent of the multiverse analysis, did not cover the entirety of the true multiverse. As mentioned in the manuscript, it was already known that the whole Self-Transcendence scale was associated with [^{11}C]SCH23390 BP_{ND} in this sample based on the results of a previous masters thesis which examined the relationship of BP_{ND} with all scales of the TCI. This study was performed independently of the original study: we only examined the TCI-DI scale in this particular investigation, with selection of items from this scale performed blind to the results of the previous study, and the image analysis and quantification were performed using more advanced tools and different regions of interest. However, it is nonetheless highly probable that both the questionnaire and the binding potential values tested in this study are fairly highly correlated to the values tested before.

This is an important consideration, and especially for expensive methodologies such as PET: exploratory investigations can diminish the inductive "value" or "potential" of a data set, even for subsequent studies. As described in the introduction, a p value is essentially a measure of surprise given an assumption of the null hypothesis being true. Surprising outcomes become less surprising with more opportunities to exhibit unusual behaviour (i.e. error rates increase), and correction for multiple comparisons is an important safeguard against this. Similarly, multiple analyses of the data as it is collected to decide whether to halt or continue data collection must be corrected for using sequential analysis methods as this may also increase the error rate[287]. However, when different investigators examine the same data from different

perspectives, it is not entirely clear how these multiple comparisons should be corrected for. Previous comparisons of similar, but not identical, data do allow investigators some insight into which comparisons will yield the most "surprising" outcomes in subsequent studies. Caution is therefore warranted in embracing exploratory research practices, since it decreases the subsequent inductive value of a data set for later hypothesis-driven research.

4.9 Study IX: Dopamine D1 Receptor Availability in First-Episode Neuroleptic-Naive Psychosis Patients

As discussed in Study VIII, the inconsistency in the results of PET studies investigating the D1R in schizophrenia may be due to various confounders, among which the most important is thought to be medication. This is due to *in vitro* findings showing that antipsychotic medication may decrease levels of D1R in the prefrontal cortex[133,134]. For this reason, we suggest that it is important to study differences in D1R availability either at early stages of the illness, before antipsychotic treatment, or prior to disease onset. While we studied the latter question in Study VIII, in Study IX, we aimed to investigate the former

In this study, we compared [^{11}C]SCH23390 BP_{ND} measured in the DLPFC and striatum between 18 drug-naive first-episode psychosis patients and 17 healthy controls. This constitutes the largest sample size of drug-naive patients examined for D1R availability using PET. Based on the previous results in the literature, we hypothesised that the patients should exhibit higher D1R availability in the DLPFC compared to controls. We used both classical frequentist testing as well as Bayesian hypothesis testing to compare the groups. For the latter, we made use of informative priors for the effects of age on [^{11}C]SCH23390 BP_{ND} based on previous studies, and used zero-centred regularising priors over the effects of patient status. Specifically, for the regularising priors, we used the expected effect size as the standard deviation of half-normal distributions for the hypotheses of higher, and lower levels of D1R availability in patients compared to controls. In this way, we could compare both of these two hypotheses with the null hypothesis as well as with one another.

The collection of this data was performed over 14 years, and hence there was some degree of heterogeneity in the measured data itself. A central part of this analysis was therefore testing to make sure that these differences were not acting as confounders, and accounting for the factors which might. Differences included aspects of PET acquisition (e.g. 2D and 3D acquisition, different measurement lengths, or some participants leaving the camera prematurely), PET reconstruction (filter settings), structural MR (T1- or T2-weighted sequences), the potential for

unconscious bias in manual ROI delineation, data storage (different file formats representing data in 2 or 3 dimensions and may impact quantification), as well as biological factors (age and sex of participants). For testing these factors, we made use of equivalence testing[68]. Almost all factors were confirmed not to have large effects on BP_{ND} values. PET acquisition in 2D or 3D could not be shown not to impact the results, and so we resolved to test the association both with and without the two measurements acquired in 3D mode. While we observed a significant age × sex interaction in this data, we were able to show in two other [^{11}C]SCH23390 datasets[288,289] that this association was not significant. This suggests that the observed association was more likely to be a false positive result. Following this preliminary analysis, we defined i) calibration factors for measurements shorter than 51 minutes and removed frames from longer measurements, ii) a calibration factor for one measurement with low molar activity and high injected mass, iii) the statistical model as a multiple regression with [^{11}C]SCH23390 BP_{ND} and age and patient status as independent variables.

In contrast to our hypothesis, we observed a statistically significantly lower [^{11}C]SCH23390 BP_{ND} in the DLPFC of psychosis patients compared to healthy controls (Figure 4.8). The null hypothesis could not be rejected in the striatum. In the Bayesian analysis, we found that the data was 3.7 times more likely under the decrease hypothesis compared to the null hypothesis, indicating moderate evidence for decreases in the DLPFC. However, the data was substantially more consistent with the decrease hypothesis than they were with the increase hypothesis (over 50 times). For the striatum, the null hypothesis was favoured over both of the other models (by 3 and 8 times for the decrease and increase hypotheses respectively). We can therefore conclude that we observe a regional effect since we show moderate evidence of an effect in the DLPFC, and moderate evidence of no effect in the striatum. We also showed that exclusion of the measurements collected in 3D did not alter the outcomes.

A significant caveat of this study, and indeed of any PET studies investigating D1R availability in the cortex, is that a non-negligible proportion of the binding is attributable to 5-HT_{2A} receptor binding; approximately 25% of cortical [^{11}C]SCH23390 or [^{11}C]NNC112 binding. While some studies have suggested that 5-HT_{2A} receptors may be decreased in schizophrenia patients compared to controls[290,291], the literature has not been consistent[292–295]. However, if the observed decreases were to be entirely explained by decreased 5-HT_{2A} receptor binding, their magnitude would imply that the availability of 5-HT_{2A} would need to be decreased by 50%, which is much larger than the reported decreases. We therefore conclude that this is unlikely to fully explain the observed lower BP_{ND} in patients.

Importantly, despite the result being significant in the frequentist analysis and the hypothesis supported by the Bayesian analyses, the magnitude of the result is not so compelling. An

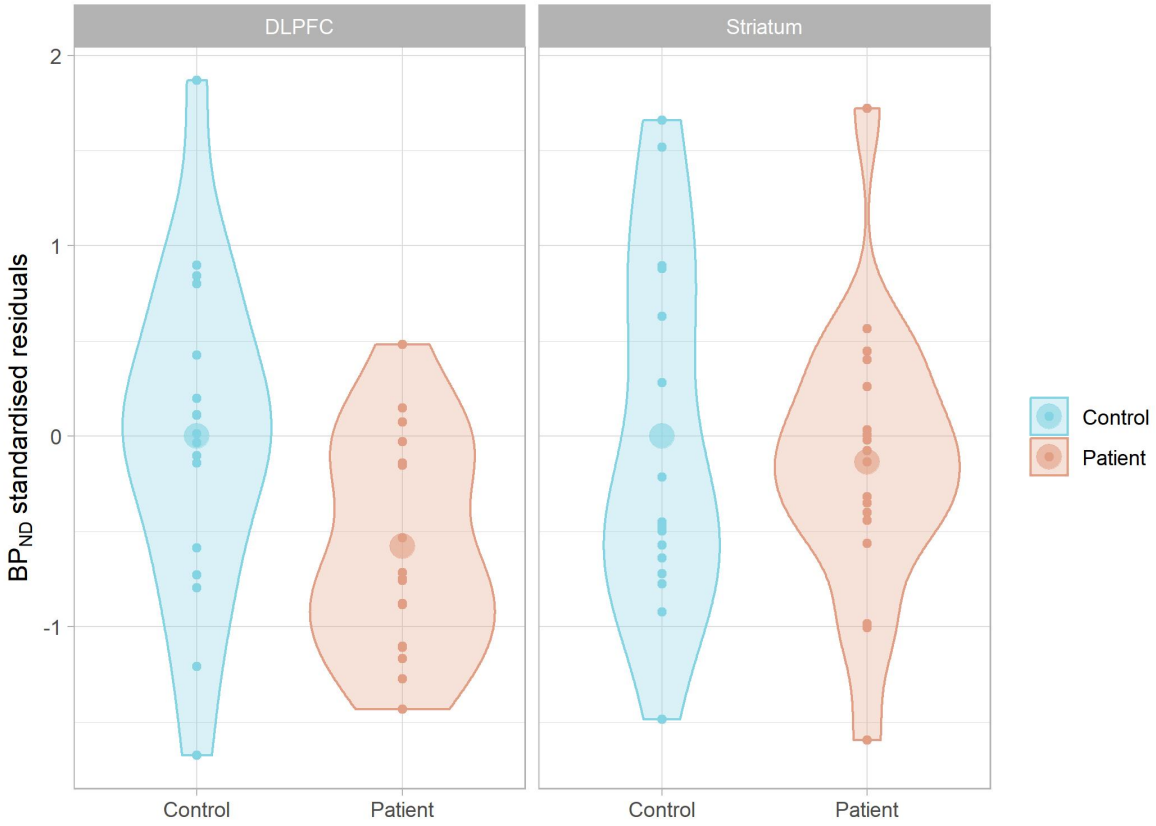

Figure 4.8: Standardised residuals representing the difference between healthy controls and psychosis patients after correction for the effects of age. Significant differences were observed in the dorsolateral prefrontal cortex (DLPFC).

analysis of the effect size in the frequentist analysis reveals that the effect size is centred over a medium effect size[277], but that the 95% confidence interval surrounding this effect size ranges from very small to very large[269]. Similarly, the effect in the striatum, despite not being significant, has a confidence interval which encompasses both a large negative effect as well as moderate positive effect. From the perspective of the Bayesian analysis, an interpretation of the BF of 3.7 as posterior model odds means that the study provides insufficient evidence to persuade someone who believed *a priori* (i.e. prior model odds) that the null hypothesis was 4 or more times more likely, to qualitatively change their beliefs.

Themes

Replicability

This study set out to replicate previous findings of differences in D1R binding between psychosis patients and controls. Based on our interpretation of literature, we considered the results of Abi-Dargham et al. (2012)[130] to be most compelling due to their comparatively large sample size, and to their sample most closely resembling our own sample, with a group of drug-naive patients. In this sense, we failed to replicate these findings, observing an effect in the opposite direction, and obtaining strong evidence that the decrease hypothesis is more likely given our data.

Reproducibility

All modelling and analysis was performed in a reproducible report. When the article is submitted, all analysis code will be made available online to accompany the article for transparency.

Transparency Statement

It is worth noting that a previous exploratory analysis of this data was conducted following collection of the data, but without the calibration factors or the confounder analyses, as well as without motion correction performed in this study. The previous analysis yielded no significant results, and a parametric analysis revealed potential decreases in the anterior cingulate cortex at an inappropriately high alpha level. The present analysis was performed completely blind to the results of the previous analysis, and we do not consider the previous results to have influenced this analysis.

Chapter 5

Future Perspectives

PET is an powerful methodology capable of deriving insights not possible with other *in vivo* imaging techniques. It is also, however, a very expensive method which additionally requires exposure of participants to radioactivity and sometimes also to arterial cannulation. It is therefore of both scientific and ethical importance that PET measurements are utilised to their full potential. This is, unfortunately, not currently the case. Most studies using PET are small: while some research questions only require very small sample sizes, many important questions for understanding the neurochemical basis of especially psychiatric disorders will require ever larger samples for deriving meaningful conclusions to advance our understanding. Another issue is that of clinical studies making use of tools, methods or outcome measures which have not been sufficiently evaluated. Both of these issues, in combination with poor application of statistical testing procedures, give rise to mixed results in many fields of clinical PET research, resulting in poor replicability and generalisability of findings. Of couse, this is not unique to PET research: failure to replicate is a substantial problem in biomedical science as a whole[58,59], and questionable research practices[79,86] and underpowered studies[78] are common across different fields of study. However, in recent years, a number of potential solutions have been gaining in popularity, primarily centred around an increasing the openness and transparency of research.

While assessment of the replicability of findings is the gold standard by which the veracity of conclusions can be evaluated, computational reproducibility serves much of the same purpose, by allowing others to see all steps taken and to determine their degree of reasonableness and correctness, but at a fraction of the cost. Additionally, when the data is provided alongside the code, other researchers are even able to assess the sensitivity of reported findings to differences in how the data may have been processed[101]. This serves as a safeguard against potentially

harmful errors[107,109].

Transparency at the level of the final analysis, however, may not capture the full degree of analytical flexibility, whether or not it may have been conscious. An important safeguard is that of pre-registration, namely recording how the data will be processed as well as the hypothesis being tested and how this will be achieved[84]. In this way, researchers can both stand by the hypotheses made prior to data collection, as well as be conscious of which hypotheses arose from observation of the data, what Nosek et al. (2018)[84] refers to as postdiction (as opposed to prediction). Of course, studies may still be conducted in a sub-optimal manner, but it still increases the statistical trustworthiness of claims made by knowing what was originally planned.

An extension of this idea is that of Registered Reports: that scientific manuscripts are submitted to journals consisting only of the introduction and methods, before any data has been collected[296]. This allows reviewers to critique the methods of a study prior to its inception, and thereby both provide helpful feedback to authors thereby limiting the potential for mistakes to be made during study design, but also to assess the scientific importance of the research *question* rather than the results. If the article is accepted, the journal makes a commitment to publish the final paper provided that the study is conducted as planned and that the conclusions are reasonable. This is assessed during a second round of peer review during which the results and discussion sections, containing the results of the pre-registered analyses as well as any additional unregistered analysis, are evaluated. This thereby reduces the positive bias observed in the literature since the incentives for authors change from producing the most compelling story to producing the most accurate one[296–298]. This further protects both authors and reviewers from bias on the part of the other. At the time of writing, there are currently 141 journals offering Registered Reports as a publishing format (https://cos.io/rr/). While of course there is also a need for exploratory research, shifting the balance in favour of confirmatory research is an important part of improving the replicability of research and thereby the generalisability of findings[85].

Reducing the degree of bias in the literature and improving the replicability of individual research findings is an excellent start, but there remains the issue that PET research is so expensive that sufficiently large samples to answer research questions in a robust manner are rare. Psychiatric PET research is therefore likely to be particularly affected by low statistical power as described in the *Introduction* section. For example, in Study IX, despite our examining the largest single sample of drug-naive schizophrenia patients, the precision of the effect size estimates were insufficient to determine whether the effect was very large or very small (i.e. 58% compared to 98% group overlap). The only solution to this widespread

issue is that of collaboration between research groups and data sharing. An example is Study VII, for which all of the individual studies did not even have 50% power to detect the final meta-analytic effect size, but when the data was combined, the conclusions were clear. The field of genetics openly embraced this strategy, where open data sharing and consortia have become central tenets, and this has yielded great dividends[299]. The neuroimaging community more broadly has begun to move in this direction with OpenfMRI[300], now renamed to OpenNeuro[301], functioning as platforms for sharing data.

Data sharing and collaboration through these platforms have also given rise to a new standard structure for sharing neuroimaging data, called the Brain Imaging Data Structure (BIDS)[103]. With a unified structure for storage of neuroimaging data in place, this has led to the development of BIDS Apps[302]. These are standardised neuroimaging pipelines which can digest and 'understand' the content of a data set, and thereby process this data without requiring substantial setup or adjustment. They do not even require installation, and run across all operating systems, due to their operating within Docker containers[303]. This also dramatically simplifies the issue of data pooling, since generalisability is diminished by the wide variety of different processing pipelines used to analyise data between research groups[102,104,302]. Re-processing multiple datasets in a homogeneous manner is rendered a substantially less demanding task when all data is stored and processed in the same manner. This is also a good real-life example of the theoretical benefits which are often proposed to result from an increased focus on reproducible research practices[106]. While PET may be lagging behind other neuroimaging methods in its adoption of these principles, the first PET BIDS datasets have been made available on OpenNeuro, and the first PET BIDS app, APPIAN, was recently released[304]. The BIDS and BIDS app framework should also pave the way for more advanced methods for image analysis and quantification of PET data to be more accessible for more research groups which lack software engineering expertise, and will simplify reproducible reporting of research findings.

In summary, I believe these initiatives to increase the transparency and openness of biomedical science can be of tremendous benefit for the research community. Greater adoption of these practices will hopefully improve the replicability of clinical PET research, but also accelerate the development, application and transmission of new and improved research methods.

Acknowledgements

I am so enormously endebted to everyone who has made it possible to reach the end of this PhD journey, both during, and before it even started. I am so thankful that I have had the opportunity to meet, get to know, and work with so many talented and wonderful people. And thank you to everyone for bearing with my being quite so awfully long-winded!

The first word of thanks goes to my principal supervisor, *Simon Cervenka*. This whole journey began with an email sent out of the blue from a masters student in the Netherlands in 2011, and what a journey it's been! It means so much to me that you have been so focused on ensuring that this whole process was first and foremost a learning experience. Your supervision has been second to none: you have given me the freedom to develop in all sorts of ways, and you have consistently had good judgement for where to draw the line when things started veering a little bit too far afield. Thank you for being so consistently supportive through every step along the way, and for your patience and kindness, which never wavered, even after our arguing for a half hour about the formulation of a single sentence.

Pontus Plavén-Sigray - we were friends since day one, and things have only gone from strength to strength since. Thank you for being my closest colleague and closest friend for these last years; and for really breaking down the difference between the two. Work and play were never far apart with you around. In addition to your major contributions to my social life these last years, I also owe an enormous part of my personal scientific development to your company, to our agreements and disagreements, to all of your input (whether or not it was requested, or even desired at the time). I couldn't have asked for a better partner with whom to graduate from the front to the back of the bus.

To my co-supervisor *Lars Farde* for always putting things into perspective: you always had the ability not only to focus on the details, but also to see everything from so far above as to see not only the bigger picture, but all the other nearby big pictures at the same time. I have really appreciated your wisdom, enthusiasm and encouragement over the years. And to my co-supervisor, *Predrag Petrovic*, thank you for your boundless enthusiasm. Your can-do

attitude and the intense discussions and brainstorming sessions that you cultivated were highly motivating, exciting and enjoyable. It is no wonder that your collaborations run far and wide, with your ability to stimulate scientific creativity in everyone around you. *Andrea Varrone*: having your stamp of approval on my work has always been a confidence-boost, and your deep reading and challenging questions, even as a middle author, have always been thoroughly appreciated. And thank you for your constant guidance, support and encouragement with all the technical developments over the years: it has always been a source of motivation.

Thank you to everyone from 1300.5: *Emma "Ottolenghi4life" Veldman*, *Jonas "Skogen" Svensson*, *Ämma "With the Ä" Tangén* and *Max "The Parmesan Man" Andersson*. Thanks for being loads of fun to hang out with, supportive friends, and for having saintly levels of tolerance and patience to put up with Pontus and I loudly arguing about everything under the sun. It's been so awesome to be surrounded with such a great bunch of people, and to be able to share all of our successes, failures and everything in-between. Thank you to the extended lunch gang, *Anton Forsberg Morén*, *Miklós Toth*, *Lenke Tari*, *Mikael Tiger* and *Vera Kerstens*, *Mahabuba Jahan* and *Jenny Häggkvist* for making the days so much more fun with a bright point in the middle, and thanks for forcing me out of the room sometimes when I couldn't drag myself away from the computer. The number of hilarious, or interesting discussions had over lunch breaks, which have continued over emails, or over days or weeks, has been testament to what a great group of people I've been surrounded by these years.

Thank you to the fun, jolly and collaborative, *Björn Schiffler*, *Lieke "Big L" de Boer*, *Nina Becker* and *William "Craig" Hedley Thompson*. Thank you all for your friendship, for all the deep conversations that were had, and for some of the most fond memories of my PhD. I've had a blast with you all at the journal clubs, Bayes club, the Brain Science pubs, and our after-works. And it has been a tremendous privilege and honour to be a part of a group who all helped each other to make such significant strides in our own personal development. FCP and JCP made a huge influence on me on so many levels, pushing me to learn things that now form a central part of how I work as a (data) scientist, teaching me about how incredible collaboration can actually feel when it *really* works, and for opening up my eyes to what was possible as "just" a group of PhD students.

Thank you to *Martin "Honeybadger don't care" Schain* firstly for patiently sitting with me as a master student when I got stuck with MATLAB, and sending me long emails of our command history: this may have been the defining factor in my finding my coding feet, which has been a gift that has kept on giving. And thank you for your friendship along the way - there has never been a dull moment with you around. And I know not to talk about Vancouver here in case there's another "poster incident". To *Patrik "Patrissimo" Fazio*: no need for meditation

with you around - your presence alone has always been enough to bring calm and warmth, and your actual words always brought even more of the same. And I've really appreciated your suggestions and company at all the jazz concerts over the last years. To *Pavitra Kannan*, for making me feel a part of this group from the start, and for being the perfect person to vent with, and for conversations that left me aching from laughing so much.

Thank you to *Karin Collste* and *Pauliina Victorsson*, the SA to the TB! I had such fun with our little choir, and I've really appreciated your guidance and help over the years on the clinical sides of my projects. And arranging for me to actually come to the clinic has been an experience which really influenced me greatly, and which I still think of frequently. There has always been a warm glow emanating from your room when you've been around!

Zsolt Cselényi and *Katarina Varnäs*: your deep technical experience has been a goldmine over the last years, and I have been so appreciative whenever I have had the opportunity to come upstairs and pick your brains. Your help and guidance have been indispensable! *Urban Hansson* and *Göran Rosenqvist*: I have always enjoyed any reason to come and knock on your door, and there has always been a laugh lying in wait. Thank you both for being so helpful and going the extra mile when helping me to understand the finer details of Linux, computer networking and PET physics, and it's been so much fun to be able to indulge in discussing all the other various technical goodies that have come up along the way.

Karin Zahir: thank you for bringing order to chaos, and calm to my panic-storm when it came to my administrative woes. And *Nina Knave*: I have always appreciated your ability to elicit real concrete change within the group. It has always been reassuring to have you on my side on any matter, as this would always mean that most of the battle was already won.

Per Stenkrona: thank you for always helping out, never saying no, and always making things work over the years. *Gina Griffioen*: although you weren't technically a student, I still had a great time supervising your project in my capacity - your motivation and enthusiastic engagement with everything made for a really rewarding collaboration. *Vladimir Stepanov* - thank you for all the invitations to join for kayaking and cycling trips. I had an awesome time every time I did come along, and thank you for saving me from the icy waters on our December excursion. That's a story that feels very fitting to have experienced with you! Thank you to *Sangram Nag* and *Mohammad Mahdi Moein* for the the tours of what goes on downstairs. These have been extremely helpful for seeing everything in context, and I really appreciate your both taking the time to help me to learn about the aspects of my PhD education which are mostly hidden from my day-to-day experiences. *Elin Johansson* and *Anders Ihrfors*: thanks for the help with all my questions over the years, and thank you for being so patient with my being so clueless with everything related to my employment etc. And

thank you for showing me the Golden Shower artwork, which still makes me laugh every time I think about it. *Camilla Gustafsson* - thanks for the company when we were two little lone masters students hidden away in the basement with the light that didn't work. Thanks to *Nalle Knutsson*, for being so welcoming and so accommodating, and for really starting off my Swedish story in style!

An enormous thank you to all the present and former members of the PET group for all the help, suggestions, laughs, reasons to eat cake, company and good work that I was able to benefit from either directly or indirectly: *Christer Halldin, Balázs Gulyás, Magdalena Nord, Aurelija Jučaite, Kai-Chun Yang, Patrik Mattson, Johan Lundberg, Ryosuke Arakawa, Jacqueline Borg, Sjoerd Finnema, Akihiro Takano, Carsten Steiger, Patricia Miranda-Azpiazu, Camilla Gustafsson, Rafael Maior, Magnus Schou, Karin Olsson, Peter Johnström, Zhisheng Jia, Maria Johansson, Mikhail Kondrashov, Prodip Datta, Youssef EL Khoury, Anton Lindberg, Arsalan Amir, Julio Gabriel, Sara Lundqvist, Susanna Nevala, Zsolt Sarnyai, Matteo Ferrante, Nadja Hellsing, Marcello Venzi, Malena Kjellén, Jonas Ahlgren, Opokua Britton-Cavaco, Åsa Södergren, Siv Eriksson, Marie Svedberg and Guennadi Jogolev.* And thank you to those whom I've inevitably missed too. And *Rasmus Berggren*, I'll include you here since you're now a part of the fika group anyhow.

And thank you to the non-KI-PET people who've also been involved in this journey too: *Vincent Millischer* for the company and our discussions that inevitably went on too long, and often involved chattering teeth in the cold after choir, *Rita Almeida* for being integral to the journal clubs, for being essential for my learning through all the seminars and tutorials, and for your help and suggestions along the way with a whole smörgåsbord of topics, *Todd Ogden* and *Francesca Zanderigo* for hosting me for a brilliant week in New York, *Anaïs Louzolo*, for the discussions, and all the stories, as well as *Diana Muessgens, Pär Flodin, Sofia Martinsson, Alva Appelgren, Gustav Nilssone, Jonathan Berrebi, Peter Fransson, Rouslan Sitnikov, Mimmi Wernman, Benjamin Garzon, Marc Guitart-Masip, Philip Pärnemets, Ida Selbing, Nathalie Wrobel, Frida Bayard, Christoph Abé, Orestis Floros, Lars Tigerström, Torbjörn Vestberg, Alexander Lebedev, David Horwath, Milena & Lars Ivansson* and *Miia Karen.*

Thanks to the *maneki-neko*: you didn't fix the lights, and you didn't even do very much beckoning, but you were a great mascot (and a great story). To the *Gaggia espresso machine and grinder*: thanks for providing the fuel for the papers of this thesis. And to Pontus' *habenula pictures*, for showing that even those things in life which appear to be most temporary can in fact be permanent.

Thank you to *Silvie* for being the best sister, whose insults always struck the perfect balance of lovingly intended, stingingly true and uproariously funny. Whenever we've had the opportunity

to catch up, I've carried the warm glow with me for days. Thank you for always grounding me by telling me how boring my work is, and reminding me how old I should feel - in the best way possible. And hi to *Steve* too: if Silvie doesn't read the thesis, and doesn't realise that she's acknowledged, please do me a favour and wait long enough before telling her so that she feels terrible for not having at least read the acknowledgements.

To *Lexx*: thank you for bringing so much joy, laughter, thought, warmth, depth, and even a bit of girth to my life. Your unique intelligence and wit are sorely missed, and the world is a lesser place without you in it.

Thank you to my friends from the Netherlands, et al: *Renate* & *Dyan*, *Theo* & *Ana*, *Axel* & *Inga*, *Iris*, *Selim*, *Michele* and *Ricardo*. And to my South African friends, *Vuma*, *Luke*, *Joe*, *Shingi*, *Tecla*. Sorry for being so horrible at keeping in touch. It always feels like there was never a break when we do speak, and you all create a diffuse feeling of home that stretches wherever you are and wherever I am in the world, and it means the world to me.

Thank you to my family back in South Africa: while it may not always be so warm over here in Sweden as it is there, I've always felt a little bit of that warmth emanating through my Skype whenever we've spoken. Thank you for everything leading up to this point, for the incredible education that you've given the opportunity to receive, for supporting me in my studies abroad, for coming all the way over to visit, and for your constant encouragement, support and love you've provided me with throughout. I could never have even begun to get here without you.

To my dearest *Dina*, thank you for turning Stockholm into not just a city, but into my home. Thank you for your companionship, your being so much fun to be around, and your always listening to me whenever I've needed to vent or rant or share successes or failures, joy and sadness (or at least for doing a good enough job of pretending to). Thank you for being the person that I'm so happy to leave work at work for. And thank you for your allowing me to work late these last few weeks to get this thesis done and dusted. I look forward to being able to repay the favour. To little *Chloé* and *Audrey*: thank you for bringing such joy into my life. I'm so happy that you're here with us, and I'm so excited to be a part of your futures! Words fail to adequately convey my feelings for my little family, but it feels like such a privilege to be a part of it.

References

1. Sejnowski, T. J., Churchland, P. S. & Movshon, J. A. Putting big data to good use in neuroscience. (2014). doi:10.1038/nn.3839

2. Wise, R. G. in *Translational neuroimaging* 1–22 (2013). doi:10.1016/B978-0-12-386945-6.00001-9

3. Morris, E. *et al.* Diagnostic accuracy of18F amyloid PET tracers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *European Journal of Nuclear Medicine and Molecular Imaging* (2016). doi:10.1007/s00259-015-3228-x

4. Bergström, M., Grahnén, A. & Långström, B. Positron emission tomography microdosing: A new concept with application in tracer and early clinical drug development. **59,** 357–366 (2003).

5. Takano, A. *et al.* Guidelines to PET measurements of the target occupancy in the brain for drug development. *European Journal of Nuclear Medicine and Molecular Imaging* **43,** 2255–2262 (2016).

6. Nordstrom, A.-L. *et al.* Central D2-dopamine receptor occupancy in relation to antipsychotic drug effects: A double-blind PET study of schizophrenic patients. *Biological Psychiatry* **33,** 227–235 (1993).

7. Farde, L. *et al.* Positron Emission Tomographic Analysis of Central D1-Dopamine and D2-Dopamine Receptor Occupancy in Patients Treated with Classical Neuroleptics and Clozapine - Relation to Extrapyramidal Side-Effects. *Archives of General Psychiatry* **49,** 538–544 (1992).

8. Farde, L., Plavén-Sigray, P., Borg, J. & Cervenka, S. Brain neuroreceptor density and personality traits: Towards dimensional biomarkers for psychiatric disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences* **373,** (2018).

9. Takahashi, H. *et al.* Differential Contributions of Prefrontal and Hippocampal Dopamine D1

and D2 Receptors in Human Cognitive Functions. *Journal of Neuroscience* **28,** 12032–12038 (2008).

10. Laruelle, M. & Huang, Y. Vulnerability of positron emission tomography radiotracers to endogenous competition. New insights. *The quarterly journal of nuclear medicine : official publication of the Italian Association of Nuclear Medicine (AIMN) [and] the International Association of Radiopharmacology (IAR)* **45,** 124–38 (2001).

11. Paterson, L. M., Tyacke, R. J., Nutt, D. J. & Knudsen, G. M. Measuring endogenous 5-HT release by emission tomography: promises and pitfalls. *J Cereb Blood Flow Metab* **30,** 1682–1706 (2010).

12. Finnema, S. J. *et al.* Application of cross-species PET imaging to assess neurotransmitter release in brain. *Psychopharmacology* **232,** 4129–4157 (2015).

13. Koepp, M. J. *et al.* Evidence for striatal dopamine release during a video game. *Nature* **393,** 266–268 (1998).

14. Dewey, S. L. *et al.* Striatal binding of the PET ligand 11C-raclopride is altered by drugs that modify synaptic dopamine levels. *Synapse* **13,** 350–356 (1993).

15. Abi-Dargham, A. *et al.* Increased baseline occupancy of D2 receptors by dopamine in schizophrenia. *Proceedings of the National Academy of Sciences* **97,** 8104–8109 (2000).

16. Fisher, B. E. *et al.* Treadmill exercise elevates striatal dopamine D2 receptor binding potential in patients with early Parkinson's disease. *NeuroReport* **24,** 509–514 (2013).

17. Haahr, M. E. *et al.* Central 5-HT4receptor binding as biomarker of serotonergic tonus in humans: A [11C]SB207145 PET study. *Molecular Psychiatry* **19,** 427–432 (2014).

18. Rinne, J. O. *et al.* Decrease in human striatal dopamine D2 receptor density with age: a PET study with [11C]raclopride. *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism* **13,** 310–4 (1993).

19. Wang, Y. *et al.* Age-dependent decline of dopamine D1 receptors in human brain: a PET study. *Synapse (New York, N.Y.)* **30,** 56–61 (1998).

20. Nord, M. *et al.* Distinct regional age effects on [11C]AZ10419369 binding to 5-HT1B receptors in the human brain. *NeuroImage* **103,** 303–308 (2014).

21. Howes, O. D. *et al.* The nature of dopamine dysfunction in schizophrenia and what this

means for treatment. *Archives of general psychiatry* **69,** 776–86 (2012).

22. Kambeitz, J. P. & Howes, O. D. The serotonin transporter in depression: Meta-analysis of in vivo and post mortem findings and implications for understanding and treating depression. *Journal of Affective Disorders* **186,** 358–366 (2015).

23. Pagano, G., Niccolini, F., Fusar-Poli, P. & Politis, M. Serotonin transporter in Parkinson's disease: A meta-analysis of positron emission tomography studies. **81,** 171–180 (2017).

24. Plavén-Sigray, P. *et al.* Positron Emission Tomography Studies of the Glial Cell Marker Translocator Protein in Patients With Psychosis: A Meta-analysis Using Individual Participant Data. (2018). doi:10.1016/j.biopsych.2018.02.1171

25. Kennedy, S. H. *et al.* Changes in Regional Brain Glucose Metabolism Measured With Positron Emission Tomography After Paroxetine Treatment of Major Depression. *American Journal of Psychiatry* **158,** 899–905 (2001).

26. Cervenka, S. *et al.* Changes in dopamine D2-receptor binding are associated to symptom reduction after psychotherapy in social anxiety disorder. *Translational Psychiatry* **2,** e120 (2012).

27. OpenStax College. *OpenStax College Physics.* (OpenStax CNX, 2016). doi:10.1063/1.3060616

28. Dahlbom, M., Eriksson, L., Rosenqvist, G. & Bohm, C. A study of the possibility of using multi-slice PET systems for 3D imaging. *IEEE Transactions on Nuclear Science* **36,** 1066–1071 (1989).

29. Cherry, S. R., Dahlbom, M. & Hoffman, E. J. 3d pet using a conventional multislice tomograph without septa. *Journal of Computer Assisted Tomography* (1991). doi:10.1097/00004728-199107000-00023

30. Spinks, T. J. *et al.* The effect of activity outside the direct field of view in a 3D-only whole-body positron tomograph. *Physics in Medicine and Biology* (1998). doi:10.1088/0031-9155/43/4/017

31. Varrone, A. *et al.* Advancement in PET quantification using 3D-OP-OSEM point spread function reconstruction with the HRRT. *European Journal of Nuclear Medicine and Molecular Imaging* (2009). doi:10.1007/s00259-009-1156-3

32. Schain, M. *et al.* Quantification of serotonin transporter availability with [11C]MADAM - A comparison between the ECAT HRRT and HR systems. *NeuroImage* **60,** 800–807 (2012).

33. Van Velden, F. H. P. *et al.* Comparison of 3D-OP-OSEM and 3D-FBP reconstruction

algorithms for High-Resolution Research Tomograph studies: Effects of randoms estimation methods. *Physics in Medicine and Biology* (2008). doi:10.1088/0031-9155/53/12/010

34. Jian, Y., Planeta, B. & Carson, R. E. Evaluation of bias and variance in low-count OSEM list mode reconstruction. *Physics in Medicine and Biology* (2015). doi:10.1088/0031-9155/60/1/15

35. Schiepers, C., Nuyts, J., Wu, H. M. & Verma, R. C. PET with18F-Fluoride: Effects of iterative versus filtered backprojection reconstruction on kinetic modeling. *IEEE Transactions on Nuclear Science* (1997). doi:10.1109/23.632737

36. Boellaard, R., Lingen, A. van & Lammertsma, A. A. Experimental and clinical evaluation of iterative reconstruction (OSEM) in dynamic PET: quantitative characteristics and effects on kinetic modeling. *Journal of Nuclear Medicine* (2001). doi:10.1080/15332840902942719

37. Lubberink, M., Boellaard, R., Weerdt, A. P. van der, Visser, F. C. & Lammertsma, A. a. Quantitative comparison of analytic and iterative reconstruction methods in 2- and 3-dimensional dynamic cardiac 18F-FDG PET. *Journal of Nuclear Medicine* (2004). doi:45/12/2008 [pii]

38. Brett, M., Penny, W. & Kiebel, S. in *Human brain function* 1–23 (2003). doi:10.1049/sqj.1969.0076

39. Andrade, A. *et al.* Detection of fMRI activation Using Cortical Surface Mapping. *Human brain mapping* **12,** 79–93 (2001).

40. Greve, D. N. *et al.* Cortical surface-based analysis reduces bias and variance in kinetic modeling of brain PET data. *NeuroImage* **92,** 225–236 (2014).

41. Innis, R. B. *et al.* Consensus nomenclature for in vivo imaging of reversibly binding radioligands. *J Cereb Blood Flow Metab* **27,** 1533–1539 (2007).

42. Lammertsma, A. A. & Hume, S. P. Simplified reference tissue model for PET receptor studies. *NeuroImage* **4,** 153–8 (1996).

43. Logan, J. *et al.* Graphical analysis of reversible radioligand binding from time-activity measurements applied to [N-11C-methyl]-(-)-cocaine PET studies in human subjects. *Journal of Cerebral Blood Flow and Metabolism* **10,** 740–747 (1990).

44. Logan, J. *et al.* Distribution Volume Ratios without Blood Sampling from Graphical Analysis of PET Data. *Journal of Cerebral Blood Flow & Metabolism* **16,** 834–840 (1996).

45. Slifstein, M. & Laruelle, M. Effects of statistical noise on graphic analysis of PET neuroreceptor studies. *Journal of nuclear medicine : official publication, Society of Nuclear*

Medicine (2000).

46. Gunn, R. N., Gunn, S. R. & Cunningham, V. J. Positron emission tomography compartmental models. *Journal of cerebral blood flow and metabolism* **21,** 635–652 (2001).

47. Perezgonzalez, J. D. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. **6,** (2015).

48. Christensen, R. Testing fisher, neyman, pearson, and bayes. *American Statistician* **59,** 121–126 (2005).

49. Pernet, C. Null hypothesis significance testing: a guide to commonly misunderstood concepts and recommendations for good practice. *F1000Research* **4,** 621 (2016).

50. Szucs, D. & Ioannidis, J. P. A. When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Frontiers in Human Neuroscience* **11,** (2017).

51. Morey, R. & Lakens, D. Why most of psychology is statistically unfalsifiable. (2016). doi:10.5281/zenodo.838685

52. Nickerson, R. S. Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods* **5,** 241–301 (2000).

53. Lecoutre, M. P., Poitevineau, J. & Lecoutre, B. Even statisticians are not immune to misinterpretations of null hypothesis significance tests. *International Journal of Psychology* **38,** 37–45 (2003).

54. Wasserstein, R. L. & Lazar, N. A. The ASA ' s statement on p-values : context , process , and purpose. *The American Statistician* **1305,** 0–17 (2016).

55. Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A. & Longobardi, C. Misconceptions of the p-value among Chilean and Italian academic psychologists. *Frontiers in Psychology* (2016). doi:10.3389/fpsyg.2016.01247

56. Colquhoun, D. The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science* **4,** (2017).

57. Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: How much can we rely on published data on potential drug targets? (2011). doi:10.1038/nrd3439-c1

58. Aarts, A. A. *et al.* Estimating the reproducibility of psychological science. *Science* **349,** aac4716 (2015).

59. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer

research. *Nature* **483,** 531–533 (2012).

60. Nosek, B. A. *et al.* Promoting an open research culture. (2015). doi:10.1126/science.aab2374

61. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* (2016). doi:10.1038/533452a

62. Camerer, C. F. *et al.* Evaluating replicability of laboratory experiments in economics. *Science* (2016). doi:10.1126/science.aaf0918

63. Benjamin, D. J. *et al.* Redefine statistical significance. (2018). doi:10.1038/s41562-017-0189-z

64. Lakens, D. *et al.* Justify your alpha. (2018). doi:10.1038/s41562-018-0311-x

65. McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. Abandon statistical significance. *arXiv preprint arXiv:1709.07588* (2017).

66. Finch, S., Cumming, G. & Thomason, N. Colloquium on Effect Sizes: the Roles of Editors, Textbook Authors, and the Publication Manual: Reporting of Statistical Inference in the Journal of Applied Psychology: Little Evidence of Reform. *Educational and Psychological Measurement* **61,** 181–210 (2001).

67. Schuirmann, D. J. A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* **15,** 657–680 (1987).

68. Lakens, D. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science* **8,** 355–362 (2017).

69. Lee, M. D. & Wagenmakers, E.-J. *Bayesian Cognitive Modeling: A Practical Course.* (Cambridge University Press, 2013).

70. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* (CRC Press, 2016).

71. Ly, A., Verhagen, J. & Wagenmakers, E. J. Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology* **72,** 19–32 (2016).

72. Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D. & Wagenmakers, E. J. The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin and Review* (2016). doi:10.3758/s13423-015-0947-8

73. Hoekstra, R., Morey, R. D., Rouder, J. N. & Wagenmakers, E. J. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin and Review* (2014).

doi:10.3758/s13423-013-0572-3

74. Fisher, R. A. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (1922). doi:10.1098/rsta.1922.0009

75. Dienes, Z. Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science* (2011). doi:10.1177/1745691611406920

76. Ioannidis, J. P. A. Why most published research findings are false. (2005). doi:10.1371/journal.pmed.0020124

77. Munafò, M. R. *et al.* A manifesto for reproducible science. (2017). doi:10.1038/s41562-016-0021

78. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience* **14,** 365–76 (2013).

79. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* **22,** 1359–1366 (2011).

80. Gelman, A. & Loken, E. The statistical Crisis in science. *American Scientist* (2014). doi:10.1511/2014.111.460

81. Elk, M. van *et al.* Meta-analyses are no substitute for registered replications: a skeptical perspective on religious priming. *Frontiers in Psychology* **6,** (2015).

82. Kerr, N. L. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* (1998). doi:10.1207/s15327957pspr0203_4

83. Nuzzo, R. How scientists fool themselves - And how they can stop. *Nature* (2015). doi:10.1038/526182a

84. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proceedings of the National Academy of Sciences* 201708274 (2018). doi:10.1073/pnas.1708274114

85. Wagenmakers, E. J., Wetzels, R., Borsboom, D., Maas, H. L. van der & Kievit, R. A. An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science* (2012). doi:10.1177/1745691612463078

86. John, L. K., Loewenstein, G. & Prelec, D. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science* **23,** 524–532

(2012).

87. Fraser, H., Parker, T., Nakagawa, S., Barnett, A. & Fidler, F. Questionable research practices in ecology and evolution. *PLoS ONE* (2018). doi:10.1371/journal.pone.0200303

88. Camerer, C. F. *et al.* Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* **2,** 637–644 (2018).

89. Baumgartner, R., Joshi, A., Feng, D., Zanderigo, F. & Ogden, R. T. Statistical evaluation of test-retest studies in PET brain imaging. *EJNMMI Research* **8,** 13 (2018).

90. Martin Bland, J. & Altman, D. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *The Lancet* (1986). doi:10.1016/S0140-6736(86)90837-8

91. Quan, H. & Shih, W. J. Assessing reproducibility by the within-subject coefficient of variation with random effects models. *Biometrics* (1996). doi:10.2307/2532835

92. Weir, J. P. J. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of strength and conditioning research / National Strength & Conditioning Association* **19,** 231–40 (2005).

93. Carrasco, J. L., Caceres, A., Escaramis, G. & Jover, L. Distinguishability and agreement with continuous data. *Statistics in Medicine* **33,** 117–128 (2014).

94. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. **86,** 420–428 (1979).

95. Nunnally, J. C. *Psychometric Theory.* 701 (1978).

96. Fleiss, J. L. *Reliability of Measurement.* 1–33 (1986).

97. Cicchetti, D. V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* **6,** 284–290 (1994).

98. Portney, L. G. & Watkins, M. P. *Foundations of clinical research: applications to practice.* (FA Davis, 2015).

99. Simons, D. J. The Value of Direct Replication. *Perspectives on Psychological Science*

(2014). doi:10.1177/1745691613514755

100. Whitaker, K. A how to guide to reproducible research. (2018). doi:10.6084/m9.figshare.5886475.v1

101. Peng, R. D. Reproducible research in computational science. (2011). doi:10.1126/science.1213847

102. Carp, J. The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage* (2012). doi:10.1016/j.neuroimage.2012.07.004

103. Gorgolewski, K. J. *et al.* The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data* (2016). doi:10.1038/sdata.2016.44

104. Nørgaard, M. *et al.* Cerebral serotonin transporter measurements with [11C]DASB: A review on acquisition and preprocessing across 21 PET centres. (2018). doi:10.1177/0271678X18770107

105. Peng, R. D., Dominici, F. & Zeger, S. L. Reproducible epidemiologic research. *American Journal of Epidemiology* **163,** 783–789 (2006).

106. Sandve, G. K. *et al.* Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology* (2013). doi:10.1371/journal.pcbi.1003285

107. Baggerly, K. A. & Coombes, K. R. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics* (2009). doi:10.1214/09-AOAS291

108. Ince, D. The Duke University scandal - what can be done? (2011). doi:10.1111/j.1740-9713.2011.00505.x

109. Reinhart, C. M. & Rogoff, K. S. Growth in a Time of Debt. *American Economic Review* **100,** 573–578 (2010).

110. Krugman, P. The Excel Depression. *The New York Times* 2013–2015 (2013).

111. Herndon, T., Ash, M. & Pollin, R. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge journal of economics* **38,** 257–279 (2014).

112. McGrath, J., Saha, S., Chant, D. & Welham, J. Schizophrenia: A concise overview of incidence, prevalence, and mortality. **30,** 67–76 (2008).

113. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait - Evidence from a meta- analysis of twin studies. *Archives of General Psychiatry* **60,** 1187–1192 (2003).

114. Fletcher, P. C. & Frith, C. D. Perceiving is believing: a Bayesian approach to explaining

the positive symptoms of schizophrenia. *Nature Reviews Neuroscience* **10,** 48–58 (2009).

115. Creese, I., Burt, D. R. & Snyder, S. H. Dopamine receptor binding predicts clinical and pharmacological potencies of antischizophrenic drugs. *Science (New York, N.Y.)* **192,** 481–3 (1976).

116. Howes, O. D. & Kapur, S. The dopamine hypothesis of schizophrenia: Version III - The final common pathway. *Schizophrenia Bulletin* **35,** 549–562 (2009).

117. Kapur, S. Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry* **160,** 13–23 (2003).

118. Slifstein, M. *et al.* Deficits in Prefrontal Cortical and Extrastriatal Dopamine Release in Schizophrenia: A Positron Emission Tomographic Functional Magnetic Resonance Imaging Study. *JAMA psychiatry* **72,** 316–24 (2015).

119. Howes, O. D. *et al.* Elevated striatal dopamine function linked to prodromal signs of schizophrenia. *Arch Gen Psychiatry* **66,** 13–20 (2009).

120. Howes, O. D. *et al.* Dopamine synthesis capacity before onset of psychosis: A prospective [18F]-DOPA PET imaging study. *American Journal of Psychiatry* **168,** 1311–1317 (2011).

121. Howes, O. D. *et al.* Progressive increase in striatal dopamine synthesis capacity as patients develop psychosis: a PET study. *Molecular Psychiatry* **16,** 885–886 (2011).

122. Hall, H. *et al.* Distribution of D1- and D2-dopamine receptors, and dopamine and its metabolites in the human brain. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* **11,** 245–256 (1994).

123. Selemon, L. D. & Goldman-Rakic, P. S. The reduced neuropil hypothesis: A circuit based model of schizophrenia. **45,** 17–25 (1999).

124. Callicott, J. H. Physiological Dysfunction of the Dorsolateral Prefrontal Cortex in Schizophrenia Revisited. *Cerebral Cortex* **10,** 1078–1092 (2000).

125. Wagstyl, K. *et al.* Multiple markers of cortical morphology reveal evidence of supragranular thinning in schizophrenia. *Translational Psychiatry* **6,** e780 (2016).

126. Okubo, Y. *et al.* Decreased prefrontal dopamine D1 receptors in schizophrenia revealed by PET. *Nature* **385,** 634–636 (1997).

127. Abi-Dargham, A. *et al.* Prefrontal Dopamine D 1 Receptors and Working Memory in

Schizophrenia. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **22,** 3708–3719 (2002).

128. Karlsson, P., Farde, L., Halldin, C. & Sedvall, G. PET study of D(1) dopamine receptor binding in neuroleptic-naive patients with schizophrenia. *The American journal of psychiatry* **159,** 761–767 (2002).

129. Kosaka, J. *et al.* Decreased binding of [11C]NNC112 and [11C]SCH23390 in patients with chronic schizophrenia. *Life Sciences* **86,** 814–818 (2010).

130. Abi-Dargham, A. *et al.* Increased prefrontal cortical D1 receptors in drug naive patients with schizophrenia: a PET study with [11C]NNC112. *J Psychopharmacol* **26,** DOI: 10.1177/0269881111409265 (2012).

131. Poels, E. M. P., Girgis, R. R., Thompson, J. L., Slifstein, M. & Abi-Dargham, A. In vivo binding of the dopamine-1 receptor PET tracers [11C]NNC112 and [11C]SCH23390: A comparison study in individuals with schizophrenia. *Psychopharmacology* **228,** 167–174 (2013).

132. Hirvonen, J. *et al.* Brain dopamine d1 receptors in twins discordant for schizophrenia. *Am J Psychiatry* **163,** 1747–1753 (2006).

133. Lidow, M. S. & Goldman-Rakic, P. S. A common action of clozapine, haloperidol, and remoxipride on D1- and D2-dopaminergic receptors in the primate cerebral cortex. *Proceedings of the National Academy of Sciences of the United States of America* **91,** 4353–4356 (1994).

134. Lidow, M. S., Elsworth, J. D. & Goldman-Rakic, P. S. Down-regulation of the D1 and D5 dopamine receptors in the primate prefrontal cortex by chronic treatment with antipsychotic drugs. *The Journal of pharmacology and experimental therapeutics* **281,** 597–603 (1997).

135. Knable, M. B., Hyde, T. M., Murray, A. M., Herman, M. M. & Kleinman, J. E. A postmortem study of frontal cortical dopamine D1 receptors in schizophrenics, psychiatric controls, and normal controls. *Biological Psychiatry* **40,** 1191–1199 (1996).

136. Cervenka, S. PET radioligands for the dopamine D1-receptor : application in psychiatric disorders. *Neuroscience Letters* 19–21 (2018). doi:10.1016/j.neulet.2018.03.007

137. Dameshek, W. White blood cells in dementia praecox and dementia paralytica. *Arch Neurol Psychiatry* **24,** 855 (1930).

138. Bradbury, T. N. & Miller, G. A. Season of Birth in Schizophrenia. A Review of Evidence,

Methodology, and Etiology. **98,** 569–594 (1985).

139. Mortensen, P. B. *et al.* Effects of family history and place and season of birth on the risk of schizophrenia. *The New England journal of medicine* **340,** 603–8 (1999).

140. Arias, I. *et al.* Infectious agents associated with schizophrenia: A meta-analysis. *Schizophrenia Research* **136,** 128–136 (2012).

141. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).

142. Upthegrove, R., Manzanares-Teson, N. & Barnes, N. M. Cytokine function in medication-naive first episode psychosis: A systematic review and meta-analysis. *Schizophrenia Research* **155,** 101–108 (2014).

143. Müller, N. & Schwarz, M. J. The immunological basis of glutamatergic disturbance in schizophrenia: Towards an integrated view. 269–280 (2007). doi:10.1007/978-3-211-73574-9-33

144. Upthegrove, R. & Barnes, N. M. The immune system and schizophrenia: an update for clinicians. *Advances in Psychiatric Treatment* **20,** 83–91 (2014).

145. Papadopoulos, V. *et al.* Translocator protein (18 kDa): new nomenclature for the peripheral-type benzodiazepine receptor based on its structure and molecular function. *Trends in Pharmacological Sciences* **27,** 402–409 (2006).

146. Owen, D. R. J. & Matthews, P. M. Imaging Brain Microglial Activation Using Positron Emission Tomography and Translocator Protein-Specific Radioligands. *International Review of Neurobiology* (2011). doi:10.1016/B978-0-12-387718-5.00002-X

147. Liu, G. J. *et al.* The 18 kDa translocator protein, microglia and neuroinflammation. *Brain Pathology* (2014). doi:10.1111/bpa.12196

148. Venneti, S., Lopresti, B. J. & Wiley, C. A. The peripheral benzodiazepine receptor (Translocator protein 18 kDa) in microglia: From pathology to imaging. (2006). doi:10.1016/j.pneurobio.2006.10.002

149. Tóth, M. *et al.* Acute neuroinflammation in a clinically relevant focal cortical ischemic stroke model in rat: longitudinal positron emission tomography and immunofluorescent tracking. *Brain Structure and Function* (2016). doi:10.1007/s00429-014-0970-y

150. Cumming, P. *et al.* Sifting through the surfeit of neuroinflammation tracers. *Journal of*

Cerebral Blood Flow & Metabolism 0271678X1774878 (2017). doi:10.1177/0271678X17748786

151. Fujita, M. *et al.* Comparison of four11C-labeled PET ligands to quantify translocator protein 18 kDa (TSPO) in human brain: (R)-PK11195, PBR28, DPA-713, and ER176—based on recent publications that measured specific-to-non-displaceable ratios. **7,** (2017).

152. Kobayashi, M. *et al.* 11C-DPA-713 has much greater specific binding to translocator protein 18 kDa (TSPO) in human brain than11C-(R)-PK11195. *Journal of Cerebral Blood Flow and Metabolism* **38,** 393–403 (2018).

153. Jučaite, A. *et al.* Kinetic analysis and test-retest variability of the radioligand [11C](R)-PK11195 binding to TSPO in the human brain - a PET study in control subjects. *EJNMMI Research* (2012). doi:10.1186/2191-219X-2-15

154. Plavén-Sigray, P. *et al.* Test-retest reliability and convergent validity of (R)-[11C]PK11195 outcome measures without arterial input function. *bioRxiv* (2018). doi:https://doi.org/10.1101/298992

155. Plavén-Sigray, P. *et al.* Accuracy and reliability of [¹¹C]PBR28 specific binding estimated without the use of a reference region. *bioRxiv* (2018).

156. Coughlin, J. M. *et al.* Regional brain distribution of translocator protein using [11C]DPA-713 PET in individuals infected with HIV. *Journal of NeuroVirology* (2014). doi:10.1007/s13365-014-0239-5

157. Park, E. *et al.* 11C-PBR28 imaging in multiple sclerosis patients and healthy controls: test-retest reproducibility and focal visualization of active white matter areas. *European Journal of Nuclear Medicine and Molecular Imaging* (2015). doi:10.1007/s00259-015-3043-4

158. Collste, K. *et al.* Test–retest reproducibility of [11C]PBR28 binding to TSPO in healthy control subjects. *European Journal of Nuclear Medicine and Molecular Imaging* **43,** 173–183 (2016).

159. Ottoy, J. *et al.* [¹⁸ F]PBR111 PET Imaging in Healthy Controls and Schizophrenia: Test – Retest Reproducibility and Quantification of Neuroinflammation. *Journal of Nuclear Medicine* (2018). doi:10.2967/jnumed.117.203315

160. Berckel, B. N. van *et al.* Microglia Activation in Recent-Onset Schizophrenia: A Quantitative (R)-[11C]PK11195 Positron Emission Tomography Study. *Biological Psychiatry* **64,** 820–822 (2008).

161. Doorduin, J. *et al.* Neuroinflammation in Schizophrenia-Related Psychosis: A PET Study.

Journal of Nuclear Medicine **50,** 1801–1807 (2009).

162. Doef, T. F. van der *et al.* In vivo (R)-[11C]PK11195 PET imaging of 18kDa translocator protein in recent onset psychosis. *npj Schizophrenia* **2,** 16031 (2016).

163. Di Biase, M. A. *et al.* PET imaging of putative microglial activation in individuals at ultra-high risk for psychosis, recently diagnosed and chronically ill with schizophrenia. *Translational psychiatry* **7,** e1225 (2017).

164. Holmes, S. E. *et al.* In vivo imaging of brain microglial activity in antipsychotic-free and medicated schizophrenia: A [11C](R)-PK11195 positron emission tomography study. *Molecular Psychiatry* **21,** 1672–1679 (2016).

165. Takano, A. *et al.* Peripheral benzodiazepine receptors in patients with chronic schizophrenia: a PET study with [11C]DAA1106. *The international journal of neuropsychopharmacology / official scientific journal of the Collegium Internationale Neuropsychopharmacologicum (CINP)* **13,** 943–950 (2010).

166. Kenk, M. *et al.* Imaging Neuroinflammation in Gray and White Matter in Schizophrenia: An In-Vivo PET Study With [18 F]-FEPPA. *Schizophrenia Bulletin* **41,** 85–93 (2015).

167. Coughlin, J. M. *et al.* In vivo markers of inflammatory response in recent-onset schizophrenia: a combined study using [11C]DPA-713 PET and analysis of CSF and plasma. *Translational Psychiatry* **6,** e777 (2016).

168. Hafizi, S. *et al.* Imaging Microglial Activation in Untreated First-Episode Psychosis: A PET Study With [18F]FEPPA. *American Journal of Psychiatry* **174,** 118–124 (2017).

169. Bloomfield, P. S. *et al.* Microglial Activity in People at Ultra High Risk of Psychosis and in Schizophrenia: An [11 C]PBR28 PET Brain Imaging Study. *The American journal of psychiatry* **35,** 2110–2119 (2016).

170. Collste, K. *et al.* Lower levels of the glial cell marker TSPO in drug-naive first-episode psychosis patients as measured using PET and [11C]PBR28. *Molecular Psychiatry* (2017). doi:10.1038/mp.2016.247

171. Peters, E. R., Joseph, S. A. & Qarety, P. A. Measurement of Delusional Ideation in the Normal Population: Introducing the PDI (Peters et aL Delusions Inventory). *Schizophrenia Bulletin* **25,** 553–576 (1999).

172. Peters, E. R., Joseph, S., Day, S. & Garety, P. Measuring delusional ideation: the 21-item

Peters et al. Delusions Inventory (PDI). *Schizophrenia Bulletin* **30,** 1005–1022 (2004).

173. Bell, V., Halligan, P. W. & Ellis, H. D. The Cardiff Anomalous Perceptions Scale (CAPS): A new validated measure of anomalous perceptual experience. *Schizophrenia Bulletin* **32,** 366–377 (2006).

174. Schürhoff, F. *et al.* Familial aggregation of delusional proneness in schizophrenia and bipolar pedigrees. *American Journal of Psychiatry* **160,** 1313–1319 (2003).

175. Freeman, D. Delusions in the nonclinical population. *Current Psychiatry Reports* **8,** 191–204 (2006).

176. Os, J. van, Linscott, R. J., Myin-Germeys, I., Delespaul, P. & Krabbendam, L. A systematic review and meta-analysis of the psychosis continuum: evidence for a psychosis proneness-persistence-impairment model of psychotic disorder. *Psychological Medicine* **39,** 179–195 (2009).

177. Lawrie, S. M., Hall, J., McIntosh, A. M., Owens, D. G. C. & Johnstone, E. C. The 'continuum of psychosis': scientifically unproven and clinically impractical. *The British Journal of Psychiatry* **197,** 423–425 (2010).

178. Linscott, R. & Os, J. A systematic review and meta-analysis of the psychosis continuum: Epidemiological evidence on the pathway from proneness to persistence to disorder. *Schizophrenia Research* **136,** S63 (2012).

179. Lenzenweger, M. F. Thinking Clearly About Schizotypy: Hewing to the Schizophrenia Liability Core, Considering Interesting Tangents, and Avoiding Conceptual Quicksand. *Schizophrenia Bulletin* **41,** 483–491 (2015).

180. Schmack, K. *et al.* Delusions and the role of beliefs in perceptual inference. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **33,** 13701–12 (2013).

181. Teufel, C. *et al.* Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proceedings of the National Academy of Sciences* (2015). doi:10.1073/pnas.1503916112

182. Powers, A. R., Mathys, C. & Corlett, P. R. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science* (2017). doi:10.1126/science.aan3458

183. Cloninger, C. R., Svrakic, D. M. & Przybeck, T. R. A Psychobiological Model of Temperament and Character. *Archives of General Psychiatry* (1993).

doi:10.1001/archpsyc.1993.01820240059008

184. Gillespie, N. A., Cloninger, C. R., Heath, A. C. & Martin, N. G. The genetic and environmental relationship between Cloninger's dimensions of temperament and character. *Personality and Individual Differences* (2003). doi:10.1016/S0191-8869(03)00042-4

185. Guillem, F., Bicu, M., Semkovska, M. & Debruille, J. B. The dimensional symptom structure of schizophrenia and its association with temperament and character. *Schizophrenia Research* (2002). doi:10.1016/S0920-9964(01)00257-2

186. Daneluzzo, E., Stratta, P. & Rossi, A. The contribution of temperament and character to schizotypy multidimensionality. *Comprehensive Psychiatry* (2005). doi:10.1016/j.comppsych.2004.07.010

187. Calvó de Padilla, M. *et al.* Temperament traits associated with risk of schizophrenia in an indigenous population of Argentina. (2006). doi:10.1016/j.schres.2005.12.848

188. Glatt, S. J., Stone, W. S., Faraone, S. V., Seidman, L. J. & Tsuang, M. T. Psychopathology, personality traits and social development of young first-degree relatives of patients with schizophrenia. *British Journal of Psychiatry* (2006). doi:10.1192/bjp.bp.105.016998

189. Smith, M. J., Cloninger, C. R., Harms, M. P. & Csernansky, J. G. Temperament and character as schizophrenia-related endophenotypes in non-psychotic siblings. *Schizophrenia Research* (2008). doi:10.1016/j.schres.2008.06.025

190. Cortés, M. J. *et al.* Psychopathology and personality traits in psychotic patients and their first-degree relatives. *European Psychiatry* (2009). doi:10.1016/j.eurpsy.2009.06.002

191. Brown, A. K. *et al.* Radiation Dosimetry and Biodistribution in Monkey and Man of 11C-PBR28: A PET Radioligand to Image Inflammation. *Journal of Nuclear Medicine* (2007). doi:10.2967/jnumed.107.044842

192. Owen, D. R. *et al.* An 18-kDa Translocator Protein (TSPO) polymorphism explains differences in binding affinity of the PET radioligand PBR28. *Journal of Cerebral Blood Flow and Metabolism* **32,** 1–5 (2012).

193. Kreisl, W. C. *et al.* A genetic polymorphism for translocator protein 18 kDa affects both in vitro and in vivo radioligand binding in human brain to this putative biomarker of neuroinflammation. *Journal of Cerebral Blood Flow and Metabolism* (2013). doi:10.1038/jcbfm.2012.131

194. Andrée, B. *et al.* The PET radioligand [carbonyl-(11)C]desmethyl-WAY-100635 binds to

5-HT(1A) receptors and provides a higher radioactive signal than [carbonyl-(11)C]WAY-100635 in the human brain. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* (2002).

195. Cselényi, Z., Lundberg, J., Halldin, C., Farde, L. & Gulyás, B. Joint explorative analysis of neuroreceptor subsystems in the human brain: Application to receptor-transporter correlation using PET data. *Neurochemistry International* **45,** 773–781 (2004).

196. Lundberg, J., Odano, I., Olsson, H., Halldin, C. & Farde, L. Quantification of 11C-MADAM binding to the serotonin transporter in the human brain. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* **46,** 1505–1515 (2005).

197. Lundberg, J., Halldin, C. & Farde, L. Measurement of serotonin transporter binding with PET and [11C]MADAM: A test-retest reproducibility study. *Synapse* **60,** 256–263 (2006).

198. Jovanovic, H. *et al.* A PET study of 5-HT1A receptors at different phases of the menstrual cycle in women with premenstrual dysphoria. *Psychiatry Research - Neuroimaging* (2006). doi:10.1016/j.pscychresns.2006.05.002

199. Lundberg, J., Borg, J., Halldin, C. & Farde, L. A PET study on regional coexpression of 5-HT1A receptors and 5-HTT in the human brain. *Psychopharmacology* **195,** 425–433 (2007).

200. Jovanovic, H. *et al.* Sex differences in the serotonin 1A receptor and serotonin transporter binding in the human brain measured by PET. *NeuroImage* **39,** 1408–1419 (2008).

201. Stenkrona, P., Halldin, C. & Lundberg, J. 5-HTT and 5-HT1A receptor occupancy of the novel substance vortioxetine (Lu AA21004). A PET study in control subjects. *European Neuropsychopharmacology* **23,** 1190–1198 (2013).

202. Nord, M., Finnema, S. J., Halldin, C. & Farde, L. Effect of a single dose of escitalopram on serotonin concentration in the non-human and human primate brain. *The international journal of neuropsychopharmacology / official scientific journal of the Collegium Internationale Neuropsychopharmacologicum (CINP)* **16,** 1577–86 (2013).

203. Nord, M., Finnema, S. J., Schain, M., Halldin, C. & Farde, L. Test-retest reliability of [11C]AZ10419369 binding to 5-HT 1B receptors in human brain. *European Journal of Nuclear Medicine and Molecular Imaging* **41,** 301–307 (2014).

204. Roland, P. E. *et al.* Human brain atlas: For high-resolution functional and anatomical mapping. *Human Brain Mapping* **1,** 173–184 (1994).

205. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical Surface-Based Analysis: I. Segmentation

and Surface Reconstruction. *NeuroImage* **9,** 179–194 (1999).

206. Fischl, B. & Dale, A. M. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 11050–11055 (2000).

207. Fischl, B., Sereno, M. I., Tootell, R. B. H. & Dale, A. M. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping* **8,** 272–284 (1999).

208. Fischl, B., Sereno, M. I. & Dale, A. M. Cortical Surface-Based Analysis: II: Inflation, Flattening, and a Surface-Based Coordinate System. *NeuroImage* **9,** 195–207 (1999).

209. Schain, M. *et al.* Improved mapping and quantification of serotonin transporter availability in the human brainstem with the HRRT. *European Journal of Nuclear Medicine and Molecular Imaging* **40,** 228–237 (2013).

210. Halldin, C. *et al.* Preparation of 11C-labelled SCH 23390 for the in vivo study of dopamine D-1 receptors using positron emission tomography. *International Journal of Radiation Applications & Instrumentation Part A, Applied Radiation & Isotopes* **37,** 1039–43 [PMID: 3027000] (1986).

211. Ekelund, J. *et al.* In vivo DA D1 receptor selectivity of NNC 112 and SCH 23390. *Molecular Imaging and Biology* **9,** 117–125 (2007).

212. Briard, E. *et al.* Synthesis and evaluation of two candidate 11C-labeled radioligands for brain peripheral benzodiazepine receptors. *Journal of Labelled Compounds and Radiopharmaceuticals* **48,** S71 (2005).

213. Chen, M. K., Baidoo, K., Verina, T. & Guilarte, T. R. Peripheral benzodiazepine receptor imaging in CNS demyelination: Functional implications of anatomical and cellular localization. *Brain* (2004). doi:10.1093/brain/awh161

214. Hall, H. *et al.* Autoradiographic localization of 5-HT1A receptors in the post-mortem human brain using [3H]WAY-100635 and [11C]WAY-100635. *Brain Research* **745,** 96–108 (1997).

215. Hirvonen, J. *et al.* Measurement of serotonin 5-HT1A receptor binding using positron emission tomography and [carbonyl-(11)C]WAY-100635-considerations on the validity of cerebellum as a reference region. *Journal of cerebral blood flow and metabolism : official*

journal of the International Society of Cerebral Blood Flow and Metabolism **27,** 185–95 (2007).

216. Shrestha, S. *et al.* Serotonin-1A receptors in major depression quantified using PET: Controversies, confounds, and recommendations. *NeuroImage* **59,** 3243–3251 (2012).

217. Halldin, C. *et al.* [11C]MADAM, a new serotonin transporter radioligand characterized in the monkey brain by PET. *Synapse* **58,** 173–183 (2005).

218. Wienhard, K. *et al.* The ECAT EXACT HR: performance of a new high resolution positron scanner. *Journal of Computer Assisted Tomography* **18,** 110–118 (1994).

219. Jong, H. W. A. M. de *et al.* Performance evaluation of the ECAT HRRT: an LSO-LYSO double layer high resolution, high sensitivity scanner. *Physics in Medicine and Biology* (2007). doi:10.1088/0031-9155/52/5/019

220. Cselényi, Z., Olsson, H., Farde, L. & Gulyás, B. Wavelet-Aided Parametric Mapping of Cerebral Dopamine D 2 Receptors Using the High Affinity PET Radioligand FLB 457. *Neuroimage* **17,** 47–60 (2002).

221. Cselényi, Z. *et al.* A comparison of recent parametric neuroreceptor mapping approaches based on measurements with the high affinity PET radioligands [11C]FLB 457 and [11C]WAY 100635. *NeuroImage* **32,** 1690–1708 (2006).

222. Turkheimer, F. E., Aston, J. A. D., Banati, R. B., Riddell, C. & Cunningham, V. J. A linear wavelet filter for parametric imaging with dynamic PET. *IEEE Transactions on Medical Imaging* **22,** 289–301 (2003).

223. Brändström, S. *et al.* Swedish normative data on personality using the Temperament and Character Inventory. *Comprehensive Psychiatry* (1998). doi:10.1016/S0010-440X(98)90070-0

224. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* (2016). doi:10.1177/1745691616658637

225. JASP Team. JASP (Version 0.9)[Computer software]. (2018).

226. Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* (2003). doi:10.1.1.13.3406

227. Carpenter, B. *et al.* Stan: A probabilistic programming language. *Journal of Statistical Software* (2017). doi:10.18637/jss.v076.i01

228. Bürkner, P.-C. brms : An R Package for Bayesian Multilevel Models Using Stan. *Journal*

of Statistical Software (2017). doi:10.18637/jss.v080.i01

229. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2018).

230. Dickey, J. M. The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics* (1971). doi:10.2307/2958475

231. Dienes, Z. Using Bayes to get the most out of non-significant results. **5,** 781 (2014).

232. Verhagen, J. & Wagenmakers, E. J. Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General* (2014). doi:10.1037/a0036731

233. Wagenmakers, E. J., Verhagen, J. & Ly, A. How to quantify the evidence for the absence of a correlation. *Behavior Research Methods* **48,** 413–426 (2016).

234. Ly, A., Etz, A., Marsman, M. & Wagenmakers, E. J. Replication Bayes factors from evidence updating. (2018). doi:10.3758/s13428-018-1092-x

235. Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel*. (Cambridge University Press, 2007). doi:10.1017/CBO9781107415324.004

236. McElreath, R. Multilevel Regression as Default. (2017).

237. Hirvonen, J., Nagren, K., Kajander, J. & Hietala, J. Measurement of cortical dopamine D1 receptor binding with 11C[SCH23390]: A test-retest analysis. *Journal of Cerebral Blood Flow & Metabolism* **21,** 1146–1150 (2001).

238. Kaller, S. *et al.* Test–retest measurements of dopamine D1-type receptors using simultaneous PET/MRI imaging. *European Journal of Nuclear Medicine and Molecular Imaging* (2017). doi:10.1007/s00259-017-3645-0

239. Lyoo, C. H. *et al.* Cerebellum Can Serve As a Pseudo-Reference Region in Alzheimer Disease to Detect Neuroinflammation Measured with PET Radioligand Binding to Translocator Protein. *Journal of Nuclear Medicine* **56,** 701–706 (2015).

240. Nair, A. *et al.* Test-retest analysis of a non-invasive method of quantifying [11C]-PBR28 binding in Alzheimer's disease. *EJNMMI Research* (2016). doi:10.1186/s13550-016-0226-3

241. Kreisl, W. C. *et al.* Stroke incidentally identified using improved positron emission tomography for microglial activation. *Archives of neurology* **66,** 1288–9 (2009).

242. Nair, A. *et al.* Erratum to: Test-retest analysis of a non-invasive method of quantifying [11C]-PBR28 binding in Alzheimer's disease (EJNMMI Research, (2016), 6, 1, (72),

10.1186/s13550-016-0226-3). (2017). doi:10.1186/s13550-017-0256-5

243. Reppert, S. M. & Weaver, D. R. Coordination of circadian timing in mammals. **418,** 935–941 (2002).

244. VanderLeest, H. T. *et al.* Seasonal Encoding by the Circadian Pacemaker of the SCN. *Current Biology* **17,** 468–473 (2007).

245. Rosenthal, N. *et al.* Seasonal affective disorder: A description of the syndrome and preliminary findings with light therapy. *Archives of General Psychiatry* **41,** 72–80 (1984).

246. Levitan, R. D. The chronobiology and neurobiology of winter seasonal affective disorder. *Dialogues in Clinical Neuroscience* **9,** 315–324 (2007).

247. Wirz-Justice, A. Biological rhythm disturbances in mood disorders. *International clinical psychopharmacology* **21 Suppl 1,** S11–S15 (2006).

248. Li, J. Z. *et al.* Circadian Patterns of Gene Expression in The Human Brain and Disruption in Major Depressive Disorder. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 9950–9955 (2013).

249. Golden, R. N. *et al.* The efficacy of light therapy in the treatment of mood disorders: A review and meta-analysis of the evidence. *American Journal of Psychiatry* **162,** 656–662 (2005).

250. Hickie, I. B. & Rogers, N. L. Novel melatonin-based therapies: Potential advances in the treatment of major depression. *The Lancet* **378,** 621–631 (2011).

251. Lewy, A. J., Lefler, B. J., Emens, J. S. & Bauer, V. K. The circadian basis of winter depression. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 7414–7419 (2006).

252. Wesemann, W., Rotsch, M., Schulz, E., Sturm, G. & Zöfel, P. Circadian rhythm of serotonin binding in rat brain–I. Effect of the light-dark cycle. *Chronobiology international* **3,** 135–9 (1986).

253. Birkett, M. & Fite, K. V. Diurnal variation in serotonin immunoreactivity in the dorsal raphe nucleus. *Brain Research* **1034,** 180–184 (2005).

254. Bucht, G., Adolfsson, R., Gottfries, C. G., Roos, B. E. & Winblad, B. Distribution of 5-hydroxytryptamine and 5-hydroxyindoleacetic acid in human brain in relation to age, drug influence, agonal status and circadian variation. *Journal of Neural Transmission* **51,** 185–203

(1981).

255. Carlsson, A., Svennerholm, L. & Winblad, B. Seasonal and Circadian Monoamine Variations in Human Brains Examined Post Mortem. *Acta Psychiatr Scand* **SUPPL 280,** 75–83 (1980).

256. Azmitia, E. C. Evolution of Serotonin : Sunlight to Suicide. *Handbook of Behavioral Neurobiology of Serotonin* **21,** 3–22 (2010).

257. Akiyoshi, J., Kuranaga, H., Tsuchiyama, K. & Nagayama, H. Circadian rhythm of serotonin receptor in rat brain. *Pharmacology, Biochemistry and Behavior* **32,** 491–493 (1989).

258. Nagayama, H. & Lu, J. Q. Circadian and circannual rythms in the function of central 5-HT1A receptors in laboratory rats. *Psychopharmacology* **135,** 279–283 (1998).

259. Spindelegger, C. *et al.* Light-dependent alteration of serotonin-1A receptor binding in cortical and subcortical limbic regions in the human brain. *World Journal of Biological Psychiatry* **13,** 413–422 (2012).

260. Praschak-Rieder, N. *et al.* Seasonal Variation in Human Brain Serotonin Transporter Binding. *Archives of General Psychiatry* **65,** 1072 (2008).

261. Kalbitzer, J. *et al.* Seasonal Changes in Brain Serotonin Transporter Binding in Short Serotonin Transporter Linked Polymorphic Region-Allele Carriers but Not in Long-Allele Homozygotes. *Biological Psychiatry* **67,** 1033–1039 (2010).

262. McMahon, B. *et al.* Seasonal difference in brain serotonin transporter binding predicts symptom severity in patients with seasonal affective disorder. *Brain* **139,** 1605–1614 (2016).

263. DeLorenzo, C. *et al.* In vivo variation in same-day estimates of metabotropic glutamate receptor subtype 5 binding using [11C]ABP688 and [18F]FPEB. *Journal of Cerebral Blood Flow and Metabolism* **37,** 2716–2727 (2017).

264. Gelman, A. & Loken, E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time. *Unpublished* (2013).

265. Gelman, A. & Carlin, J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science* **9,** 641–651 (2014).

266. Vul, E., Harris, C., Winkielman, P. & Pashler, H. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition1. *Perspectives on Psychological Science*

4, 274–290 (2009).

267. Spearman, C. The proof and measurement of association between two things. *The American journal of psychology* **15,** 72–101 (1904).

268. Nunnally, J. C. Introduction to psychological measurement. (1970).

269. Sawilowsky, S. S. New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods* **8,** 597–599 (2009).

270. Hedge, C., Powell, G. & Sumner, P. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods* (2017). doi:10.3758/s13428-017-0935-1

271. Vukasović, T. & Bratko, D. Heritability of personality: A meta-analysis of behavior genetic studies. *Psychological Bulletin* **141,** 769–785 (2015).

272. Kendler, K. S., Kuhn, J. & Prescott, C. A. The Interrelationship of Neuroticism, Sex, and Stressful Life Events in the Prediction of Episodes of Major Depression. *American Journal of Psychiatry* **161,** 631–636 (2004).

273. Khan, A. A., Jacobson, K. C., Gardner, C. O., Prescott, C. A. & Kendler, K. S. Personality and comorbidity of common psychiatric disorders. *British Journal of Psychiatry* **186,** 190–196 (2005).

274. Borg, J., Andrée, B., Soderstrom, H. & Farde, L. The serotonin system and spiritual experiences. *American Journal of Psychiatry* **160,** 1965–1969 (2003).

275. Karlsson, H., Hirvonen, J., Salminen, J. K. & Hietala, J. No association between serotonin 5-HT1A receptors and spirituality among patients with major depressive disorders or healthy volunteers. *Molecular Psychiatry* **16,** 282–285 (2011).

276. Jeffreys, H. Theory of Probability. *Oxford Classic Texts in the Physical Sciences* **V,** 1–40 (1961).

277. Cohen, J. Statistical power analysis for the behavioral sciences. **2nd,** 567 (1988).

278. Owen, D. R. J. *et al.* Mixed-Affinity Binding in Humans with 18-kDa Translocator Protein Ligands. *Journal of Nuclear Medicine* **52,** 24–32 (2011).

279. Kreisl, W. C., Henter, I. D. & Innis, R. B. Imaging Translocator Protein as a Biomarker of Neuroinflammation in Dementia. (2017). doi:10.1016/bs.apha.2017.08.004

280. Jucaite, A., Forssberg, H., Karlsson, P., Halldin, C. & Farde, L. Age-related reduction in

dopamine D1 receptors in the human brain: From late childhood to adulthood, a positron emission tomography study. *Neuroscience* **167,** 104–110 (2010).

281. Thomas, D., Radji, S. & Benedetti, A. Systematic review of methods for individual patient data meta-analysis with binary outcomes. *BMC Medical Research Methodology* **14,** 79 (2014).

282. Upthegrove, R., Manzanares-Teson, N. & Barnes, N. M. Cytokine function in medication-naive first episode psychosis: A systematic review and meta-analysis. *Schizophrenia Research* **155,** 101–108 (2014).

283. Notter, T. *et al.* Translational evaluation of translocator protein as a marker of neuroinflammation in schizophrenia. *Molecular Psychiatry* **23,** 323–334 (2018).

284. Narayan, N. *et al.* The macrophage marker translocator protein (TSPO) is down-regulated on pro-inflammatory 'M1' human macrophages. *PLoS ONE* **12,** (2017).

285. Owen, D. R. *et al.* Pro-inflammatory activation of primary microglia and macrophages increases 18 kDa translocator protein expression in rodents but not humans. *Journal of Cerebral Blood Flow and Metabolism* **37,** 2679–2690 (2017).

286. Quintana, D. S. From pre-registration to publication: A non-technical primer for conducting a meta-analysis to synthesize correlational data. **6,** (2015).

287. Lakens, D. Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology* (2014). doi:10.1002/ejsp.2023

288. Bäckman, L. *et al.* Dopamine D1 receptors and age differences in brain activation during working memory. *Neurobiology of Aging* **32,** 1849–1856 (2011).

289. De Boer, L. *et al.* Attenuation of dopamine-modulated prefrontal value signals underlies probabilistic reward learning deficits in old age. *eLife* **6,** (2017).

290. Rasmussen, H. *et al.* Decreased frontal serotonin2A receptor binding in antipsychotic-naive patients with first-episode schizophrenia. *Archives of General Psychiatry* (2010). doi:10.1001/archgenpsychiatry.2009.176

291. Rasmussen, H. *et al.* Low frontal serotonin 2A receptor binding is a state marker for schizophrenia? *European Neuropsychopharmacology* (2016). doi:10.1016/j.euroneuro.2016.04.008

292. Lewis, R. *et al.* Serotonin 5-HT2 Receptors in Schizophrenia: A PET Study Using [18F]Setoperone in Neuroleptic-Naive Patients and Normal Subjects. *American Journal of*

Psychiatry **156,** 72–78 (1999).

293. Trichard, C. *et al.* No serotonin 5-HT(2A) receptor density abnormality in the cortex of schizophrenic patients studied with PET. *Schizophrenia Research* (1998). doi:10.1016/S0920-9964(98)00014-0

294. Okubo, Y. *et al.* Serotonin 5-HT2 receptors in schizophrenic patients studied by positron emission tomography. *Life Sci* (2000).

295. Erritzoe, D. *et al.* Cortical and Subcortical 5-HT2A Receptor Binding in Neuroleptic-Naive First-Episode Schizophrenic Patients. *Neuropsychopharmacology* (2008). doi:10.1038/sj.npp.1301656

296. Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P. & Willmes, K. Registered Reports: Realigning incentives in scientific publishing. *Cortex* **66,** 1–2 (2015).

297. Scargle, J. D. Publication bias: the 'File Drawer' problem in scientific inference. *Journal of Scientific Exploration* (2000).

298. Simonsohn, U., Nelson, L. D. & Simmons, J. P. P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General* (2014). doi:10.1037/a0033242

299. Singleton, A. B. & Traynor, B. J. For complex disease genetics, collaboration drives progress. (2015). doi:10.1126/science.aaa9838

300. Poldrack, R. A. *et al.* Toward open sharing of task-based fMRI data: the OpenfMRI project. *Frontiers in Neuroinformatics* (2013). doi:10.3389/fninf.2013.00012

301. Gorgolewski, K., Esteban, O., Schaefer, G., Wandell, B. & Poldrack, R. OpenNeuro—a free online platform for sharing and analysis of neuroimaging data. *Organization for Human Brain Mapping. Vancouver, Canada* 1677 (2017).

302. Gorgolewski, K. J. *et al.* BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLOS Computational Biology* **13,** e1005209 (2017).

303. Boettiger, C. An Introduction to Docker for Reproducible Research. *SIGOPS Oper. Syst. Rev.* **49,** 71–79 (2015).

304. Funck, T., Larcher, K., Toussaint, P.-J., Evans, A. C. & Thiel, A. APPIAN: Automated Pipeline for PET Image Analysis. **12,** 64 (2018).