Impact of analytical variability on data compatibility in functional Magnetic Resonance Imaging studies

Xavier Rolland

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1
COMUE UNIVERSITÉ BRETAGNE LOIRE

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Xavier ROLLAND

Impact of Analytical Variability on Data Compatibility in Functional Magnetic Resonance Imaging Studies

Thèse présentée et soutenue à Rennes, le 16/05/2022
Unité de recherche : Empenn

Direction de la Thèse :

Dir. de thèse :	Christian Barillot	Directeur de Recherche	CNRS
Co-dir. de thèse :	Pierre Maurel	Maître de Conférence	Université de Rennes 1
Encadrante :	Camille Maumet	Chargée de Recherche	Inria

Rapporteurs avant soutenance :

Michel Dojat	Directeur de Recherche	Inserm
Daniel Margulies	Directeur de Recherche	CNRS

Composition du Jury :

Président :	Jean-Marc Jézéquel	Professeur	Université de Rennes 1
Examinateurs :	Sorina Caramasu Pop	Ingénieure de Recherche	CNRS
	Michel Dojat	Directeur de Recherche	Inserm
	Daniel Margulies	Directeur de Recherche	CNRS
	Pierre Maurel	Maître de Conférence	Université de Rennes 1
	Camille Maumet	Chargée de Recherche	Inria

Acknowledgements

I would like to start by thanking my PhD supervisors: Camille Maumet and Christian Barillot, with whom I started this PhD thesis, and Pierre Maurel, who kindly agreed to become supervisor of the PhD thesis after Christian. Their experience, knowledge and advices, as well as their availability and support, have always been a precious help for me during these years as a PhD student in the Empenn team.

I would also like to thank the members of my PhD supervision committee, Mathieu Acher and Elisa Fromont, for their important feedbacks and remarks each year during the CSI meetings.

Many thanks also to all the members of jury: Michel Dojat and Daniel Margulies for the review of my PhD thesis; Jean-Marc Jézéquel for agreeing to be president of the jury for my PhD defense; and all of them, along with Sorina Pop, for agreeing to be part of the jury and for their many relevant comments regarding my work during my PhD defense.

Many thanks also to my friends and family for their support during these years. I would in particular like to thank my coworkers with whom I shared many victories at the Synthi quizz and played many games of darts and billiards, as well as online games during the lockdown in 2020 and 2021.

Finally, I would like to thank all of the Empenn team for the great time spent together during my years at Empenn. It was a pleasure to work here and I hope that we will have the chance to meet again often.

Xavier ROLLAND

Contents

Résumé en français **ix**
0.1 Contexte . . . ix
0.1.1 Imagerie cérébrale et IRM fonctionnelle . . . ix
0.1.2 Chaînes de Traitement en IRM fonctionnelle . . . x
0.1.3 Reproductibilité . . . xii
0.2 Estimation et Correction de l'Effet de la Variabilité Analytique . . . xii
0.2.1 Compatibilité entre les données . . . xii
0.2.2 Variabilité Analytique . . . xiii
0.2.3 Estimation de la Validité des Analyses de Groupe Combinant des Données de Sujets Traitées Différemment . . . xiv
0.2.4 Correction de la Variabilité Analytique Dans les Analyses de Groupe Combinant des Données de Sujets Traités Différemment . . . xv

Introduction **1**

I Context **4**

1 Brain Imaging and Functional Magnetic Resonance Imaging **5**
1.1 Brain Imaging . . . 5
1.1.1 The Early Days of Brain Imaging . . . 5
1.1.2 Examples of Brain Imaging Techniques . . . 7
1.2 Functional MRI . . . 8
1.2.1 Principle of MRI and BOLD fMRI . . . 8
1.2.2 Experimental Design of Task-Based BOLD fMRI Studies . . . 9
1.2.3 Data Processing and Software Packages . . . 10

2 fMRI Pipelines **11**
2.1 Preprocessing . . . 11

2.1.1 Distortion Correction . . . 12
2.1.2 Motion Correction . . . 13
2.1.3 Slice-Timing Correction . . . 14
2.1.4 Registration to a Template Space . . . 15
2.1.5 Smoothing . . . 15
2.2 First-Level Analysis . . . 17
2.2.1 General Linear Model . . . 18
2.2.2 BOLD Signal and GLM for First-Level Analysis . . . 20
2.2.3 HRF modeling . . . 25
2.3 Second-Level Analysis . . . 26
2.3.1 GLM in Second-Level Analysis . . . 26
2.3.2 Modeling of Variance in Second-Level Analysis . . . 27
2.4 Statistical Inference . . . 27
2.4.1 Hypothesis Testing . . . 28
2.4.2 Inference in fMRI studies . . . 29

3 Reproducibility **32**
3.1 Reproducibility . . . 32
3.1.1 Why Is Reproducibility Important in Scientific Research . . . 32
3.1.2 Reproducibility Crisis . . . 33
3.1.3 Causes of the Reproducibility Crisis . . . 34
3.1.4 Solutions . . . 37
3.2 Reproducibility in Neuroimaging . . . 39
3.2.1 The Reproducibility Crisis in Neuroimaging . . . 39
3.2.2 Solutions in Neuroimaging . . . 41
3.3 Conclusion . . . 44

II Contribution **46**

4 Data Compatibility **47**
4.1 Sample Sizes and Statistical Power . . . 47
4.2 Increasing Statistical Power: Combining Shared Data . . . 49
4.2.1 Data Sharing in Neuroimaging: History and State of the Art . . . 49
4.2.2 Re-use of Shared Data and Combination of Data . . . 50
4.3 A New Challenge: How To Assess The Compatibility of Data Processed Differently? 51
4.3.1 Data Combination and Subject-Level Processing . . . 51

4.3.2 Research Question: What Is The Impact of Analytical Variability On Data Compatibility? . . . 52

5 Analytical Variability 54
5.1 Types of Variability . . . 54
5.1.1 Inter-Subject Variability . . . 54
5.1.2 Test-Retest Variability . . . 55
5.1.3 Inter-Scanner and Inter-Site Variability . . . 55
5.1.4 Analytical Variability . . . 56
5.2 Analytical Variability in fMRI . . . 56
5.2.1 Why does Analytical Variability Exist? . . . 56
5.2.2 Variability in Pipeline Parameters . . . 57
5.2.3 Variability in Software Conditions . . . 60
5.3 Research Questions related to Analytical Variability . . . 62
5.3.1 Variability in the final results . . . 62
5.3.2 Analytical Flexibility . . . 64
5.3.3 Validity of the Final Results . . . 65
5.4 My Work . . . 65

6 First Contribution: Estimation of the Validity of Group Analyses Using Subject Data Processed Differently 68
6.1 Introduction . . . 68
6.2 Material and Methods . . . 71
6.2.1 Material . . . 72
6.2.2 Subject-Level Pipelines . . . 72
6.2.3 Between-group Analyses . . . 73
6.2.4 False Positive Rates Estimation . . . 74
6.3 Results . . . 75
6.3.1 Analyses using the same pipeline . . . 75
6.3.2 Between-group analyses with different HRF models . . . 75
6.3.3 Between-group analyses with different levels of smoothing . . . 80
6.3.4 Between-group analyses with different number of motion regressors . . . 80
6.3.5 Combined effects of parameters . . . 81
6.4 Discussions . . . 82
6.4.1 Conclusion . . . 85

7 Second Contribution: Correction of Pipeline Effects in Group Analyses 90
7.1 Overview of The Literature . . . 90

7.1.1 ComBat 91
7.1.2 Surrogate Variable Analysis 92
7.1.3 Remove Unwanted Variation 93
7.1.4 Applications to Neuroimaging 94
7.2 Methods 95
7.3 Results 98
7.3.1 Results Without Correction 98
7.3.2 Results With Correction 98
7.4 Discussion and Further Works 101

Conclusion **103**

Résumé en français

Cette partie constitue un résumé en français du contenu du manuscrit, qui a été rédigé en anglais.

Dans la première partie du manuscrit, je propose une mise en contexte du le travail qui a été réalisé au cours de la thèse, en présentant le domaine de la neuroimagerie, et celui de l'Imagerie par Résonance Magnétique Fonctionnelle (IRMf) en particulier, et la crise de la reproductibilité. La seconde partie décrit la question que l'on souhaite aborder, et les réponses que nous y avons apporté avec nos travaux de recherche : nous proposons de tirer parti de la large quantité de données partagées disponibles aujourd'hui en neuroimagerie pour combiner ces données, afin d'effectuer des analyses avec de plus grandes tailles d'échantillons. Cela peut conduire à faire des analyses utilisant des données pré-traitées différemment au niveau individuel. Par conséquent, nous cherchons à étudier l'effet de la variabilité liée à ces différences de traitement, appelée variabilité analytique, sur la validité des résultats d'analyses de groupes combinant des données de sujet traitées différemment.

0.1 Contexte

0.1.1 Imagerie cérébrale et IRM fonctionnelle

La neuroimagerie a été à l'origine de nombreux progrès dans le domaine des sciences cognitives, en procurant au chercheur des méthodes d'observation in vivo du cerveau, le plus souvent pas ou peu invasives. Elle permet d'étudier le cerveau sur plusieurs aspects : l'imagerie structurelle, qui étudie la morphologie et la structure du cerveau, et l'imagerie fonctionnelle, qui mesure l'activité du cerveau pour permettre d'obtenir des informations sur son fonctionnement (Figure 1).

Différentes modalités existent en neuroimagerie, afin de mesurer l'activité électrique dans le cerveau. Cette mesure peut se faire soit directement, soit indirectement, grâce notamment au lien entre le débit sanguin cérébral et l'activation des régions du cerveau. Parmi ces modalités, on trouve notamment l'électroencéphalographie (EEG) ou la Tomographie par Emission de Positons (TEP). La modalité qui nous intéresse est l'Imagerie par Résonance Magnétique (IRM), qui est

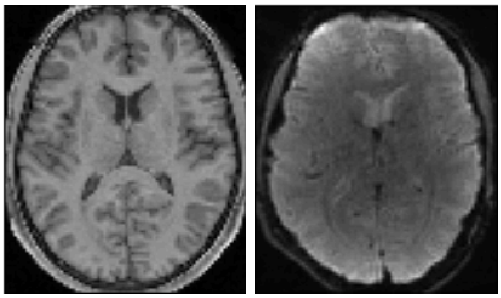


Figure 1 – Exemples d'images structurelle (gauche) et fonctionnelle (droite) obtenue sur un sujet du Human Connectome Project [152].

une méthode non-invasive d'observation du cerveau.

L'Imagerie par Résonance Magnétique Fonctionnelle (IRMf) utilise l'IRM pour acquérir une séquence temporelle d'images 3D du cerveau, afin de mesurer les variations d'un signal IRMf à chaque position du cerveau. L'une des principales méthodes d'IRMf est l'IRMf de tâche, qui étudie l'activation dans le cerveau des participants lors de l'exécution de tâches spécifiques. Il existe plusieurs techniques d'IRMf qui mesurent différentes caractéristiques associées à l'activation du cerveau. La principale technique d'IRMf est l'IRMf "Blood-Oxygen Level Dependent" (BOLD), qui estime le niveau d'activité cérébrale à l'aide du niveau d'oxygénation sanguine, qui est lié à l'afflux sanguin dans le cerveau, lui-même lié à l'activité cérébrale.

Pour étudier certaines aptitudes cognitives, un paradigme est défini pour demander aux participants d'une étude d'effectuer des tâches pendant qu'ils sont dans la machine IRM. Les images obtenues sont ensuite utilisées pour étudier la relation existant entre les variations temporelles d'activité cérébrales aux différentes zones du cerveau et les périodes ou instants de réalisation de tâches.

0.1.2 Chaînes de Traitement en IRM fonctionnelle

Les études d'IRMf BOLD de tâche se font en appliquant une succession d'étapes de traitement et d'analyse sur les données acquises, au niveau individuel puis au niveau du groupe. Ces analyses sur les données se font à l'aide de logiciels développés pour le traitement des données de neuroimagerie. La suite d'étape de traitements appliquées sur les données s'appelle une chaîne de traitement. A chaque étape, plusieurs opérations peuvent être faites, avec plusieurs choix méthodologiques possibles. La Figure 2 illustre les étapes principales d'une chaîne de traitement en IRMf.

Au niveau individuel, la première étape des chaînes de traitement est le pré-traitement des images brutes. L'objectif est de préparer les données à l'analyse statistique qui va suivre, notamment en effectuant des opérations de correction (recalage rigide, lissage pour réduire le bruit et les différences anatomiques). Une fois le pré-traitement effectué, une première étape d'analyse

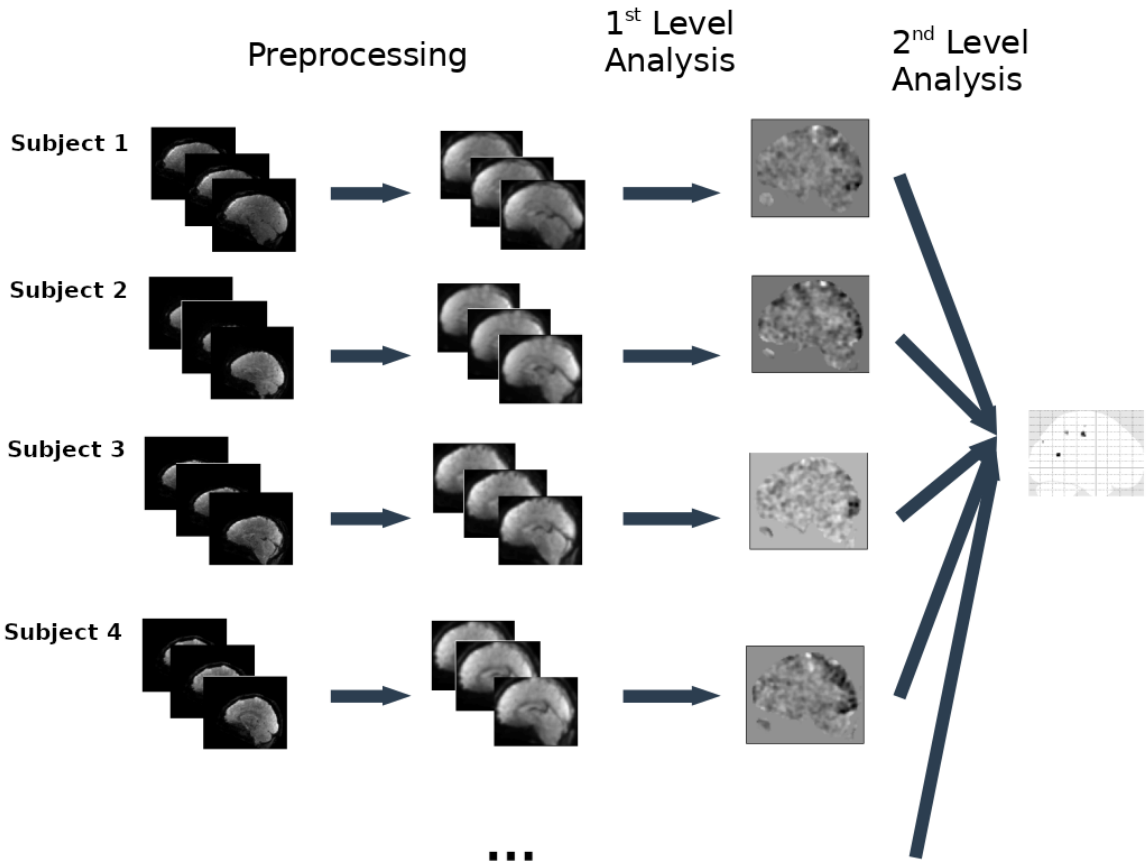


Figure 2 – Illustration des différentes étapes du traitement de données en IRMf: pré-traitement des données brutes des sujets, analyse statistique au niveau individuel donnant des cartes de contraste 3D pour chaque sujet, analyse de groupe combinant les résultats au niveau individuel, puis application de l'inférence statistique pour obtenir une carte statistique 3D seuillée.

statistique (appelée analyse au premier niveau) est appliquée au niveau individuel. Pour chaque sujet, pour chaque voxel, on dispose d'un signal temporel IRMf. Un modèle linéaire dont les régresseurs correspondent à des réponses attendues en cas d'activation pour chaque tâche réaliser, afin d'associer à chaque tâche un niveau d'activation (d'autres régresseurs peuvent également être ajoutés pour enlever du bruit dans le signal). Pour chaque sujet, une combinaison linéaire des valeurs associées à chaque régresseur (appelée contraste), correspondant à une variable d'intérêt, peut être obtenue pour chaque voxel.

Une analyse au niveau du groupe (appelée analyse au second niveau) est ensuite effectuée à l'aide des données de contrastes obtenues pour chaque sujet à la fin de l'analyse au premier niveau. Cette analyse peut permettre d'observer à chaque position du cerveau, à l'aide d'un nouveau modèle linéaire, les différences d'activation entre les sujets, pour le contraste obtenu,

en fonction d'une variable donnée : âge, appartenance à un groupe... Une carte de contraste est ensuite obtenue au second niveau, pour construire une carte statistique sur laquelle un seuillage est appliqué afin de détecter les zones du cerveau pour lesquelles on a des valeurs significatives pour la variable que l'on souhaite observer.

0.1.3 Reproductibilité

Depuis plusieurs années, dans plusieurs domaines scientifiques dont notamment la neuroimagerie, des doutes ont été émis quant à la fiabilité des résultats de recherche publiés, qui est pourtant un élément fondamental pour pouvoir mener à bien de nouvelles recherches se basant dessus. Plusieurs études ont tenté de reproduire des résultats expérimentaux dans divers domaines scientifiques, avec de faibles taux de réussite. Cela a amené la communauté scientifique à parler de "crise de la reproductibilité", pour laquelle plusieurs facteurs ont été identifiés.

Concernant la simple reproductibilité des résultats, avec des données et un protocole identique, le manque de report d'information à propos du protocole appliqué sur les données, ou l'absence de partage de donnée, comptent parmi les facteurs identifiés empêchant la reproduction des résultats. Les autres types de situations dans lesquels les chercheurs peuvent souhaiter répéter des expériences de recherche induisent la présence de source de variabilité susceptible de faire changer les résultats (différence dans le protocole ou dans les données). Dans des domaines comme la neuroimagerie, les résultats de recherche reposent sur des méthodes statistiques à partir desquelles il est possible d'obtenir des faux positifs. Une grande flexibilité méthodologique peut notamment faciliter l'obtention de faux positif. Aussi, les biais de publications peuvent entraîner une surreprésentation des faux positifs dans la littérature. Enfin, des faibles tailles d'échantillon peuvent entraîner une baisse de la puissance statistique et donc de la probabilité de détection des vrais effets, ce qui amène à une baisse de la valeur prédictive positive.

Plusieurs solutions ont été proposées pour palier à la crise de la reproductibilité, dont certaines relatives aux pratiques de recherches (incitation à la mise à disponibilité des données et du code utilisé, par exemple). En neuroimagerie, une solution concernant la faible puissance statistique est la réutilisation des données partagées pour augmenter les tailles d'échantillons.

0.2 Estimation et Correction de l'Effet de la Variabilité Analytique

0.2.1 Compatibilité entre les données

En neuroimagerie, les faibles tailles d'échantillons ont été identifié comme l'un des facteurs majeurs liés à la crise de la reproductibilité. Bien que de plus en plus d'études utilisent des grandes tailles d'échantillons et que la médiane des tailles d'échantillons augmentent, celle-ci

reste relativement faible par rapport à ce qui pourrait exister.

Une façon d'augmenter les tailles d'échantillons peut être de tirer parti du nombre croissant d'ensembles de données partagées disponibles. De nombreux projets de partage de données ont vu le jour à partir des années 2000, suite aux premiers résultats obtenus avec les initiatives pionnières, comme le fMRIDC. Ces projets comprennent notamment des études sur de larges ensembles de données rendus disponibles, ou des plate-formes permettant aux chercheurs de mettre à disposition les données qu'ils ont acquises et utilisées pour leurs études. Une façon de ré-utiliser ces données peut être de combiner les données de plusieurs ensembles de données différents. Cependant, cela peut impliquer de combiner des données traitées différemment : les ensembles de données disponibles peuvent ne contenir que des données de sujets déjà traitées, et l'on a vu que plusieurs choix méthodologiques étaient possibles dans le traitement des données.

La variabilité des résultats liée aux différents choix de traitement possibles est appelée variabilité analytique. Dans nos travaux, nous étudions la façon dont la variabilité analytique au niveau individuel a un effet sur la validité des résultats combinant des données traitées différemment. Pour ce faire, on effectue des analyses inter-groupe sous l'hypothèse nulle, en combinant des données issues de différentes chaînes de traitement, et on compare les taux de faux positifs obtenus avec ceux attendus avec le seuillage choisi.

0.2.2 Variabilité Analytique

L'étude des sources de variabilité des résultats et de leur prise en compte lors des expériences fait partie de nombreux domaines scientifiques, y compris la neuroimagerie. Parmi les sources de variabilité connues, on compte notamment la variabilité inter-sujet, la variabilité Test-Retest ou encore la variabilité technique.

La variabilité analytique est la variabilité liée aux différentes possibilités de choix méthodologiques dans les analyses. En IRMf de tâche, cela se traduit par les différentes possibilités en termes de chaînes de traitement pour une même étude. L'absence de chaîne de traitement de référence fait que plusieurs choix existent, sans qu'il soit toujours possible d'établir lesquels sont les meilleurs. De plus, les améliorations constantes des chaînes de traitement créent de nouveaux choix méthodologiques possibles. Les sources de variabilité peuvent se trouver au niveau des choix de paramètres dans les chaînes de traitement : il est possible de faire ou non certaines opérations, de les faire dans plusieurs ordres, et avec plusieurs choix de valeurs de paramètres possibles. Les conditions logicielles sont également une source de variabilité : les différents logiciels pour l'IRMf n'appliquent pas toujours par défaut les mêmes étapes de traitement (et ne peuvent pas toujours le faire), et les conditions en termes de système d'exploitation et de version des logiciels peuvent également avoir un impact sur les résultats.

Plusieurs questions de recherche peuvent alors se poser concernant la variabilité des résultats : la variabilité analytique a notamment été étudiée au niveau des résultats d'analyses de groupes

en IRMf. En comparaison, notre travail s'intéresse à l'effet de la variabilité analytique au niveau individuel, sur la validité des analyses de groupes combinant des données, comme illustré sur la Figure 3.

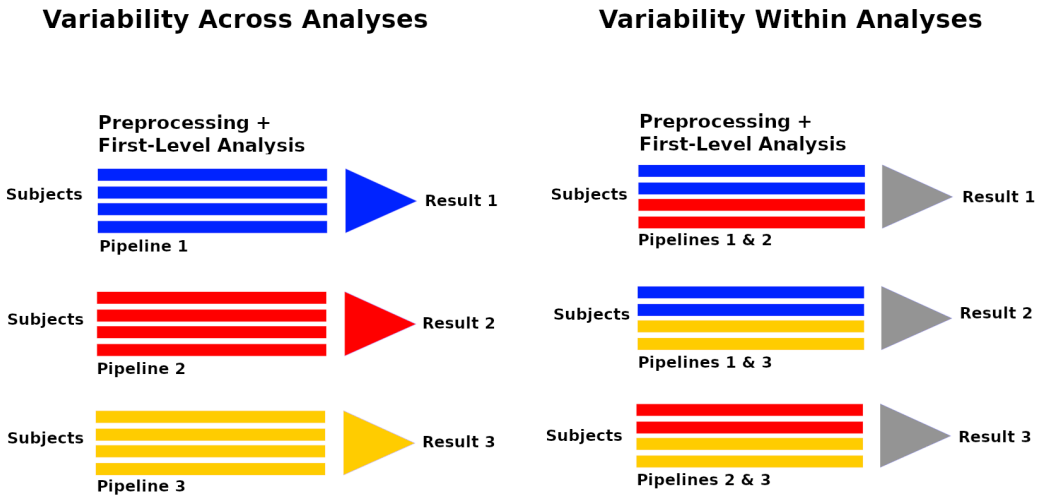


Figure 3 – Présentation des différentes situations dans lesquelles on peut étudier l'impact de la variabilité, dans des travaux existants comme [20, 14] (à gauche) et dans nos travaux (droite).

0.2.3 Estimation de la Validité des Analyses de Groupe Combinant des Données de Sujets Traitées Différemment

Notre première contribution a consisté à effectuer des analyses inter-groupes, avec deux groupes de 50 sujets, sous l'hypothèse nulle (pas de différence d'activation entre les deux groupes) pour établir la validité ou non d'une combinaison de chaînes de traitement. A chacun des deux groupes est associée une chaîne de traitement au niveau individuel, avec laquelle tous les sujets du groupe sont traités. Un seuillage corrigé pour un taux d'erreur au niveau de la famille à 0.05 est appliqué, et l'analyse de groupe est répétée 1000 fois avec des données de sujets différentes. Le taux de détection, c'est-à-dire la proportion d'analyses parmi les 1000 répétitions pour lesquelles on observe au moins une détection, sert de taux de faux positifs et est comparé au taux attendu de 0.05. Une valeur supérieure indique l'invalidité de la combinaison, car elle signifie une probabilité plus forte d'obtenir de détecter à tort un effet qui n'existe pas avec cette combinaison. Ces analyses ont été faites avec plusieurs paires de chaînes de traitement au niveau individuel, qui différaient entre elle sur trois paramètres : le lissage spatial au pré-traitement, la présence ou

l'absence de dérivées temporelles de la réponse hémodynamique dans la matrice du modèle linéaire au premier niveau, et le nombre de régresseurs de mouvement dans cette matrice. Les données utilisées pour l'étude sont les données des 1080 participants ayant réalisé les tâches pour le paradigme moteur dans l'ensemble de données 1200-subject datasets du Human Connectome Project.

Pour chaque différence de paramètre, on a pu observer les taux de faux positifs obtenus avec les comparaisons de chaînes de traitement qui différaient sur ces paramètres (Figure 4), dans les deux sens. Nos résultats permettent notamment d'observer que des différences au niveau du lissage et du nombre de régresseurs entre les chaînes de traitement avaient un effet qui causait l'invalidité des résultats. Aussi, la combinaison de plusieurs différences de paramètres entraînait la combinaison des effets, avec des taux de faux positifs plus forts notamment. Les analyses ont été faites avec deux logiciels différents, SPM et FSL. Les combinaisons intra-logiciel de chaînes de traitement donnaient, pour des comparaisons correspondantes avec les deux logiciels, des résultats similaires, comme on peut l'observer sur la Figure. D'autres résultats (P-P plots, histogramme des valeurs statistiques) pour observer la structure de la répartition des valeurs statistiques au second niveau pour ces analyses.

Nos résultats permettent de mettre en évidence le fait que, dans certains cas, la combinaison de données de sujets traitées différemment peut entraîner l'invalidité des résultats, ce qui empêche les chercheurs de combiner les données sans prendre en compte l'effet de la variabilité analytique sur les résultats.

0.2.4 Correction de la Variabilité Analytique Dans les Analyses de Groupe Combinant des Données de Sujets Traités Différemment

Notre première contribution a permis de mettre en évidence l'effet de la variabilité analytique sur la validité des analyses inter-groupes combinant des données traitées différemment. Notre seconde contribution consiste à effectuer les mêmes analyses en appliquant une méthode de correction. On estime la validité des analyses avec et sans ces méthodes de correction – et donc l'efficacité de ces méthodes – en utilisant le même cadre méthodologique que l'étude précédente, avec les mêmes variations de paramètres entre les chaînes de traitement.

Notre méthode de correction consiste à utiliser une covariable binaire associée à la chaîne de traitement utilisée pour chaque sujet, dans le modèle linéaire au second niveau. Ce type de méthode de correction est simple à mettre en œuvre, et déjà appliqué pour d'autres types de variabilité (par exemple la variabilité inter-site). Cela requiert d'avoir dans chaque groupe un mélange de données traitées avec les différentes chaînes de traitement, au lieu d'avoir des groupes distingués en fonction de la chaîne de traitement comme c'était le cas dans le travail précédent. Sinon, la différence entre les chaînes de traitements est liée à la différence entre les groupes, qui est ce que l'on souhaite observer. Pour cela, on effectue des analyses où la première chaîne de

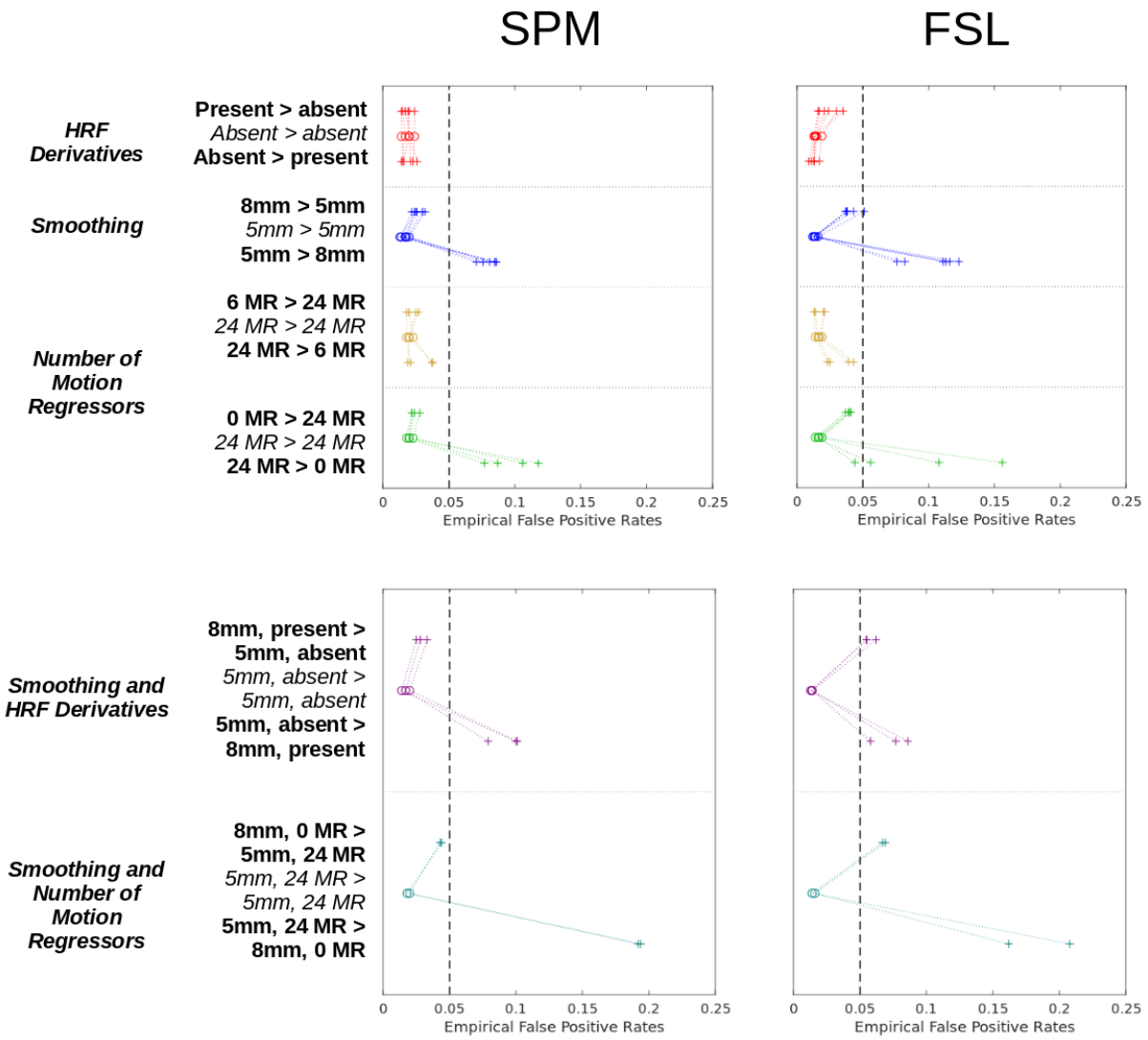


Figure 4 – Taux de faux positifs empiriques obtenus pour les analyses inter-groupes combinant deux chaînes de traitements différentes l'une de l'autre sur un (première ligne) ou deux (seconde ligne) paramètres, pour plusieurs paramètres que l'on fait varier par rapport à une valeur par défaut, dans les deux sens, sous SPM (gauche) et FSL (droite).

traitement correspond à 50%, 70%, 80% ou 90% des sujets dans le premier groupe (et autant pour la seconde chaîne de traitement dans le second groupe). Ces analyses, avec ces proportions, sont faites avec et sans les méthodes de correction appliquées pour comparer les résultats.

Sans correction appliquée, on observe qu'il n'y a pas d'effet remarquable pour 50%, c'est-à-dire lorsqu'il y a dans chaque groupe autant de données de sujets traitées avec chacune des deux chaînes de traitement. L'effet que l'on observait dans le travail précédent pour certaines différences de paramètres apparaît lorsque la proportion de la chaîne de traitement principale

dans chaque groupe augmente. En appliquant la méthode de correction, dans les cas où un effet de la variabilité est présent, celui-ci est au moins partiellement corrigé à 70 et 80% (comme on peut l'observer sur les P-P plots sur la Figure 5). En revanche, la correction ne semble pas marcher à 90%, accentuant l'effet ou ajoutant de l'invalidité là où il n'y en avait pas. Cela peut être dû au fait qu'en approchant des 100%, la différence entre les chaînes est de plus en plus proche de la différence entre les groupes. Cela peut aussi être dû au faible nombre de sujets associés à la chaîne de traitement secondaire dans chaque groupe (10% de 50 sujets = 5 sujets).

Par conséquent, il semble possible d'appliquer des méthodes permettant de corriger l'effet de la variabilité analytique dans certains cas de combinaison des données traitées différemment.

Nous avons défini dans nos travaux un cadre méthodologique qui nous a permis d'estimer la validité d'analyses inter-groupe combinant des données. Cela nous a permis de montrer que certaines différences de paramètres entre des chaînes de traitement avaient un effet pouvant entraîner l'invalidité des résultats en combinant des données traitées différemment, mais également qu'il était possible de limiter cet effet à l'aide de méthodes de correction. L'étude de l'effet de la variabilité analytique en neuroimagerie, et des manières de corriger cet effet, peut être approfondie, notamment en faisant des analyses dans des situations différentes, pour d'autres types de paradigmes, avec d'autres méthodes de correction, et également pour des modalités autres que l'IRMf BOLD de tâche. Le développement de méthodes de correction efficaces dans un grand nombre de situations permettra d'envisager la possibilité de combiner des données traitées différemment sans avoir à se soucier de l'effet de la variabilité analytique.

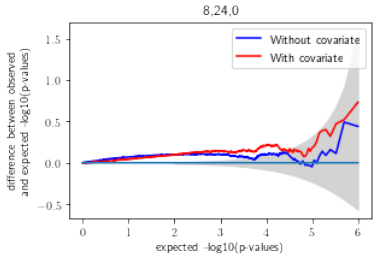
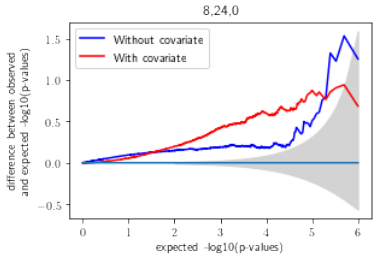
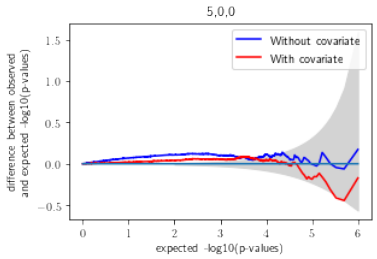
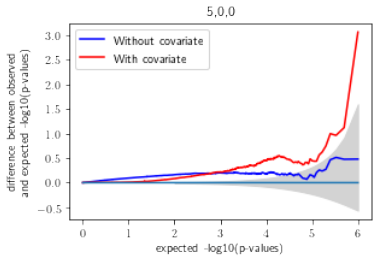
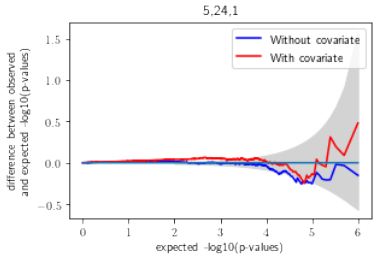
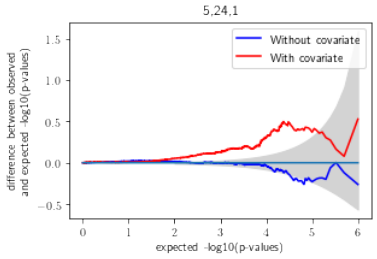
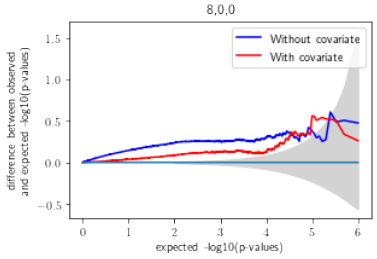
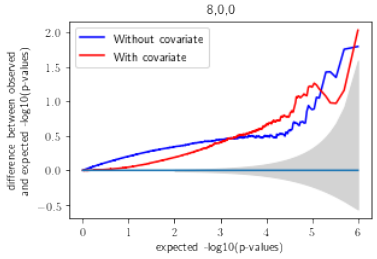
	80%	90%
Effect of Smoothing		
Effect of Regressors		
Effect of HRF		
Effect of Motion Regressors + Smoothing		

Figure 5 – Variantes de P-P plots obtenues pour des analyses inter-groupes sous SPM combinant des données de sujets traitées différemment, avec ou sans covariable pour corriger l'effet de la variabilité analytique, pour des chaînes de traitement qui diffèrent sur un ou deux paramètres. Dans chaque analyse, les données du premier groupe sont principalement traitées avec la première chaîne de traitement (80% pour les résultats dans la première colonne, 90% pour les résultats dans la seconde), et les données du second groupe principalement avec la deuxième chaîne de traitement.

Introduction

Neuroimaging techniques have played an essential part in the progresses made in cognitive neuroscience over the last decades. One of the most widely used techniques is Magnetic Resonance Imaging (MRI), which is a non-invasive method to provide different types of 3D images of the brain that can be used to study and understand the brain. More specifically, Functional Magnetic Resonance Imaging (fMRI) is an MRI modality that can be used to study brain functions. One of the main methods in fMRI is task-based fMRI, which studies activation in the brains when participants are performing specific pre-defined tasks. Various techniques, which measure different features associated to brain activation, exist in fMRI. The main fMRI technique is Blood-Oxygen level Dependent (BOLD) fMRI, which takes advantage of the relationship existing between brain activation and cerebral blood flow to estimate brain activity.

Like many other scientific fields, neuroimaging research has been subject to a reproducibility crisis. Many studies which have tried to replicate published research results, with low rates of successful replication. Since the reliability of experimental results is fundamental for the progress of scientific research, these findings have led the scientific community to question the way research results are produced. Multiple factors have been identified as potential causes of non-reproducibility. Among them, the problem of low sample sizes have been highlighted as an obstacle for the reproducibility of research results in neuroimaging. Studies in fMRI rely on statistical analyses on data from groups of participants. The probability of detecting an existing effect notably depends on the number of subjects, and low sample sizes induce low statistical power, decreasing the chance that a result detected as positive is a true positive.

Increasing sample sizes could be a solution to improve the reproducibility of research results in fMRI. To do so, researchers may take advantage of data sharing initiatives. An increasing number of datasets are made available, with data that can be re-used for new studies. Using these data can be a way to increase the number of subjects within a study and achieve larger sample sizes. In particular, researchers may consider combining in a same study available data coming from different datasets to increase sample sizes, as was done in the 1000 Functional Connectome Project for example [12].

fMRI group studies rely on the processing and statistical analysis of fMRI data, both at subject and group-level. A series of applied processing and analysis steps is called a 'pipeline'.

Each processing and analysis step contains a variety of possible choices in terms of methodology and implementation, leading to multiple possible pipelines for a same research question. This leads to a variability in results obtained at each processing level, called analytical variability, caused by differences in processing choices. To date, most analyses performed by re-using data typically runs by re-processing all datasets using the exact same data preparation steps. However, shared datasets may contain only data which have been already processed, and different datasets may use different subject-level processing pipelines.

Therefore, combining data from different datasets may require performing analyses with subject data processed differently, unlike what is usually done. Analytical variability has been studied in various situations in neuroimaging. In this thesis, we adress the question of the impact of analytical variability on results in task-based BOLD fMRI when combining data processed differently.

Here, we propose a methodological framework with which to assess the validity of group analyses using different subject-level processing pipelines. We took advantage of the theoretical properties regarding hypothesis testing used in fMRI to define criteria for the validity of analyses. Because we wanted to study the impact of variability of specific parameters in the fMRI pipelines, we performed analyses using pipelines which differed only on this set of parameters. Finally, we observed whether, for analysis where there was an effect of pipeline differences causing invalidity, applying a correction method could remove this pipeline effect, in order to be able to combine these data processed differently.

Our manuscript is organized as follows:

Part I: Context

Chapter 1: Brain Imaging and Functional MRI

This chapter provides an overview of neuroimaging, its history and its contribution to the field of cognitive neuroscience. We present some of the different existing neuroimaging modalities, techniques and methods used, before focusing our discussion on fMRI studies.

Chapter 2: fMRI Pipelines

As mentioned above, fMRI studies use pipelines to perform processing and analysis of the data in order to obtain research results. In this chapter, we present the different processing steps applied on the data, what they are used for and what are some of the possible methodological choices at each step.

Chapter 3: Reproducibility

In this chapter, we introduce the notion of reproducibility and other forms of repetitions of experimental results. We explain how it became a major concern in scientific research, leading to a reproducibility crisis, and what are the main causes that have been identified, in science in general as well as in neuroimaging in particular.

Part II: Estimation And Correction of The Effect of Analytical Variability

Chapter 4: Data Compatibility

Here, we discuss the issue of small sample sizes in neuroimaging, and the solution that we propose in order to increase them, which is to re-use shared data. We then explain what is the issue we may be confronted with in this situation: the necessity to combine data processed differently and the potential impact of analytical variability on data compatibility.

Chapter 5: Analytical Variability

After introducing analytical variability, we give details about the sources of analytical variability in neuroimaging. We give an overview of the literature on analytical variability in neuroimaging, and position our work in this context.

Chapter 6: Estimation of The Validity of Group Analyses Using Subject Data Processed Differently

Our first contribution introduces a methodological framework to assess the impact of analytical variability on the validity of between-group analyses in fMRI. We perform between-group analyses under null hypothesis, using two different pipelines, and compare the false positive rates obtained to those theoretically expected. Comparison are done with various pairs of pipelines, with each pipeline associated to one of both groups.

Chapter 7: Correction of Pipeline Effects in Group Analyses Using Subject Data Processed Differently

In this chapter, we propose a correction method on between-group analyses. Our method requires to have proportions of subjects processed by each pipeline in each group different from 100%. We thus perform between-group analyses, similarly to what we did in the previous section, but with different proportions. These analyses are performed with and without correction applied, and their validity is assessed using the same framework as in the first contribution.

Part I

Context

Chapter 1

Brain Imaging and Functional Magnetic Resonance Imaging

Task-based Functional Magnetic Resonance Imaging (fMRI) is a neuroimaging modality for the study of the relation between cognitive tasks and activation of regions in the brain. In order to introduce fMRI, in this chapter, we will explain how the development of neuroimaging techniques has provided researchers with new ways of getting information about the brain. We will describe some of these neuroimaging techniques and modalities, and how they can be used to answer various research questions in neuroscience, with a focus on fMRI.

1.1 Brain Imaging

1.1.1 The Early Days of Brain Imaging

Since their development in the 1980s, neuroimaging techniques have become an essential tool for research in cognitive neuroscience. They allow researchers to make in vivo observations of the brain, most often with low or no invasiveness. For this reason, they can be used to answer a variety of questions about the brain, about how it is structured anatomically, and what is its relationship with cognitive functions. These questions have interested scientists for a long time before the emergence of neuroimaging. However, because of the lack of non-invasive ways of making in vivo observations of the brain, possibilities to obtain information about the brain were limited.

Before the availability of brain imaging techniques, one of the main sources of information about the brain was the study of patients suffering from injuries or pathologies [138]. Observational studies were used to understand the relationship existing between cognitive dysfunctions for injured subjects and damages in specific regions of their brain. When neuroimaging techniques arrived, researchers took advantage of their potential to study the brain without the

limitations that existed previously.

There are two main types of neuroimaging techniques. Structural imaging uses neuroimaging to obtain information about brain morphology and structure (it can notably be used to detect brain lesions). On the other hand, functional imaging measures brain activity to obtain information about brain functions, for example by observing the relationship in time between cognitive processes and activation in brain regions. Figure 1.1 shows examples of structural and functional images obtained with MRI.

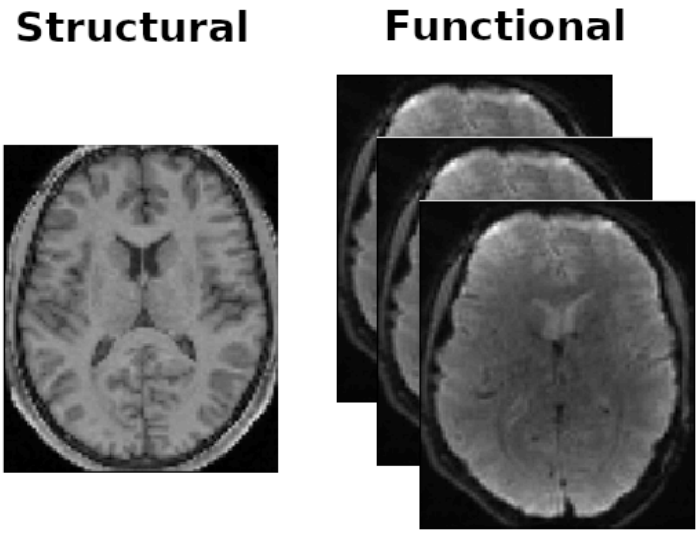


Figure 1.1 – Example of structural (left) and functional (right) images for a subject from the Human Connectome Project [152] obtained with magnetic resonance imaging. The images displayed are two-dimensional axial views of the brain, corresponding to slices from 3D images. Structural data consists in a single high-quality image of the brain while functional data contains a temporal series of images.

For a long time, the main theory regarding brain organization and its association with mental functions was the simple idea that specific brain regions were linked to specific cognitive aptitudes. This traditional conception, called localization, appeared long before the development of neuroimaging techniques. The essential argument in favor of localization was the observation that mental dysfunctions in disabled subjects were associated to physical damages in specific areas of the brain [127]. This was notably the case for Broca's area: the observation of damages in a same part of the brain for multiple subjects with reduced speech abilities led Broca to suggest that this region of the brain was associated to the speech function [34]. The development of knowledge about the brain has led researchers to question localization in favor of other models regarding the organization of brain functions. In particular, networking models account for the existing interconnection between different parts of the brain [36]. Other works have also addressed the question of brain organization through the study of gradients which represent

spatial transition within the brain regarding various functional and structural features.

Functional neuroimaging relies on the in vivo measure of brain activity. One possibility is to directly measure electrical activity within the brain. The other possibility is to take advantage of the existing relationship between neural activity and cerebral blood flow (CBF): activation of neurons in the brain leads to an increase of blood flow in the region of the brain. This association, called neurovascular coupling, has been known for more than a century [133], although the explanation behind the phenomenon is still unclear and scientific research is still active to try to better understand it [121].

Thanks to neurovascular coupling, brain activity can also be estimated indirectly through the measure of features associated to the variations in CBF. Properties about the variations in CBF associated to neuronal activity, and details about how they can be used for modelling in analyses, are given in 2.2.2.

1.1.2 Examples of Brain Imaging Techniques

The development of brain imaging techniques started in the 1980s with the use of Positron Emission Tomography (PET) [127], which allowed researchers to obtain the first insights within brain organization and localization of mental function through the observation of neurotypical individuals [129]. Since then, multiple techniques have been developed, each with their own advantages and possibilities as well as their own drawbacks. Here, we describe some of the main existing neuroimaging techniques: PET, and electroencephalography/magnetoencephalography (EEG/MEG). Since our work focuses on fMRI, details for this neuroimaging modality are given specifically in a dedicated section (section 1.2).

PET

Position Emission Tomography is based on the injection of a radioactive tracer [112]. These tracers are attached to molecules such as water or glucose, allowing to measure the variations of blood level, since blood contains the tracer. The main advantage of PET scans is the possibility to directly measure cerebral blood flow quantitatively. However, PET has several drawbacks and limitations, especially compared to other modern neuroimaging techniques: it is an invasive technique, as it requires the injection of radioactive tracers in the blood. Also, spatial resolution is low and scans require a lot of time compared to fMRI (minutes for PET scans, against seconds for fMRI).

EEG/MEG

MEG and EEG are used to measure brain activity through the electrical activity of neurons. For EEG, electrodes are positioned on the head of a subject to measure the electrical potential

at each position [13]. It allows to measure brain activity in various psychological states. Its advantages are the low cost and ease of use, as well as the very high temporal resolution. Disadvantages include the high variability in electric conductivity, and thus in measure of electric potential across subjects and within subjects over time. It also has low spatial resolution and therefore does not allow to detect precisely the localization of brain activation. Similarly, MEG measures magnetic fields at the surface of the head. MEG has the advantage of allowing a better spatial localization than EEG. however, MEG is also more difficult to use, as the equipment required to measure magnetic fields is very large and expensive [69]. Both modalities can be used in complement of each other, as well as in complement of other neuroimaging modalities such as fMRI.

1.2 Functional MRI

1.2.1 Principle of MRI and BOLD fMRI

Magnetic Resonance Imaging (MRI) aligns spinning atomic nuclei thanks to a strong magnetic fields (several Tesla). It then disturbs their axis of rotation and allows to observe the radio frequency signal generated with the nuclei returning to their baseline status [127]. MRI has the advantage of being non-invasive, having low risks and being less costly than PET scans, making it easier to do and accessible to a more important range of the population. It also has a better spatial resolution. The main drawback, compared to PET scans, is the necessity to stand still during the time of acquisition and how this reduces the possibilities in terms of experimental frameworks.

MRI can be used both for structural and functional imaging. Functional MRI data are temporal sequences of functional images of the brain obtained during an acquisition session (Fig. 1.1). Functional MRI measures variations over time of an fMRI signal at each position of the brain with an MRI machine in one or multiple subjects. This results in a temporal sequence of 3D images of the brain for each subject.

There are multiple ways of using MRI for functional neuroimaging studies. MRI can be used for the observation of basal information about the brain and assessment of functional tissue characteristics. This is the case of perfusion MRI, which can be used to obtain information about cerebral blood volume and blood flow [42]. Other than that, two main methods of fMRI exist: one is resting-state fMRI, which studies the variations of the fMRI signal and thus brain activity when subjects are not performing any task. This method can be used to observe features in the brain such as spontaneous networks of synchronous activation, and identify their differences across groups of subjects (with and without disorders for example) [92]. The other one is task-based fMRI, where subjects are asked to perform specific tasks at specific times in the MRI machine. The relation between the variation of BOLD signal and the expected responses for

each task is used to detect activity [127].

Various techniques exists using MRI to apply these functional neuroimaging methods. Arterial Spin Labeling (ASL) relies on the measure of magnetically labeled water protons as a tracer within the subject. Dynamic susceptibility contrast imaging (DSC MRI) uses gadolinium chelate injected in the blood (therefore an invasive technique). ASL and DSC MRI can be used for perfusion MRI. ASL can also be used for task-based and resting-state MRI. The most common fMRI technique is Blood-Oxygen Level Dependent (BOLD) fMRI. BOLD fMRI takes advantage of the fact that an increase in blood flow is associated to an increase in blood oxygen. The change of oxygenation leads to a variation of the signal which is measured by the MRI machine, called the fMRI signal. Therefore, the variations in the measured fMRI signal are related to brain activity [127].

1.2.2 Experimental Design of Task-Based BOLD fMRI Studies

For functional neuroimaging modalities, specific study frameworks have been developed to answer research questions using these modalities. They usually consist in acquiring neuroimaging data within subjects participating in the study, and then applying processing and analysis steps on these data to obtain results. In parallel to the development of neuroimaging research, software packages have also been created and developed to automatize the processing and analysis steps for these neuroimaging modalities. Specific details regarding the software packages used for task-based fMRI are given in 1.2.3.

BOLD fMRI can be used to estimate brain activity through the measure of variation of blood oxygenation, which is associated with variations in blood flow. The goal of task-based BOLD fMRI studies is to observe how variations in measured fMRI signal (which are associated to brain activation), at each position of the brain, are related to cognitive activities performed by the participants. Group studies can then be performed, for example to see how this brain activity varies across subjects depending on factors such as age, gender, or presence of specific mental or neurological disorders.

In practice, fMRI studies are carried on by acquiring, oten for multiple participants, a temporal sequence of MRI images of the brain [142]. While in the MRI machine, participants are asked to perform various tasks. During the session, a certain number of fMRI images are acquired at regular time interval for each subject. The time between the acquisition of two consecutive images in the sequence is called the repetition time (TR). A standard TR in task-based BOLD fMRI is usually a few seconds, with less than a second being considered a low TR [87]. The MRI session during which data is acquired can last a few minutes.

The experimental design requires to define a paradigm. This paradigm gives the guidelines regarding what the subjects will be required to do in the MRI machine, depending on the cognitive process study. A paradigm is defined by a set of tasks that must be performed and

a time-series of stimuli associated to these tasks. Various standard types of task paradigms exist to study specific cognitive abilities. In the Human Connectome Project [152], for example, data were acquired for seven paradigms: working memory, gambling, motor, language, social cognition, relational processing and emotion processing [5].

There are two main possibilities regarding the experimental designs for the tasks that can be performed within the MRI machine: block designs and event-related designs. In block designs, the stimuli last for long periods of time, whereas for event-related designs, stimuli are instantaneous and lead to short neural responses for the subjects [53]. Event-related designs appeared in neuroimaging research after block design, once knowledge about the structure of the fMRI response to neural stimuli allowed for a correct modelling of this response. An advantage of event-related designs is their greater flexibility, with the possibility of randomizing the times and order of appearance of stimuli [96]. One drawback is the low signal-to-noise ratio compared to block designs [120].

1.2.3 Data Processing and Software Packages

At the end of data acquisition, a temporal series of 3D images giving the fMRI signal intensity at each time and position of the brain is available for each subject. Processing and statistical analysis steps are applied on the data in order to obtain the research results. The analysis is done using both the brain imaging data and the stimulus time-series which give information about the tasks performed within the MRI machine for each subject.

The series of processing steps applied on the data to obtain the final results is called a pipeline [142]. As for other neuroimaging fields, studies in task-based fMRI largely rely on the automated processing of neuroimaging data. Because of this, various software packages dedicated to it have been developed since the emergence of neuroimaging. These software packages can be used to automatically perform all the steps of a pipeline.

The main neuroimaging software packages available for task-based BOLD fMRI are AFNI (Analysis of Functional Neuroimages) [29], FSL (FMRIB Software Library) [80], and SPM (Statistical Parametric Mapping) [118]. These neuroimaging software packages notably cover about 80% of task-based BOLD fMRI analyses in the scientific literature, according to [20]. Differences can exist across software packages in methods and algorithms used to perform a same operation. This variability across software packages and its consequence on research results are detailed in chapter 5.

The following chapter provides a precise description of the processing steps and possible methodological choices at each step in pipelines for task-based BOLD fMRI analyses.

Chapter 2

fMRI Pipelines

In the previous chapter, we introduced fMRI. Once subject data have been acquired, multiple steps of processing and analysis must be applied in order to obtain fMRI results. This sequence of steps for a given study is called a pipeline. First, preprocessing steps are applied on the data to clean it and prepare it for subsequent analysis. Subject-level analysis (also sometimes refered to as first-level analysis) is then performed on the data to associate a level of activation to each condition of interest at each position of the brain. This results in a 3D map of the brain for each subject and for each condition showing these levels of activation. Linear combination of these 3D maps, called contrast maps, can also be obtained. The contrast maps from multiple subjects can then be used for group studies, in a group-level analysis (also sometimes called second-level analysis): for example, to observe a difference in activation between two groups. The group-level analysis results in a single group-level 3D contrast map. Finally, thresholding can be applied on this map, based on hypothesis testing, to detect zones with significant effects. An overview of fMRI pipelines, and the data resulting from each step, can be seen on Figure 2.1.

In this chapter, we give a detailed description of each step within processing pipelines used in fMRI. We detail the different operations which can be carried out at preprocessing, first-level analysis, second-level analysis and statistical inference. At each step, there are multiple methodological possibilities which lead to a variety of possible pipelines for a same research question.

2.1 Preprocessing

Preprocessing constitutes the first set of operations applied on fMRI data. Researchers want to use the images acquired on multiple subjects to perform statistical analysis on them, first at subject-level and then at group-level. Two things prevent them from doing these analyses with the raw fMRI images. First, there is a number of artifacts in these data which have to be cleaned. Second, for group-level analysis, the positions of the voxels must correspond to the

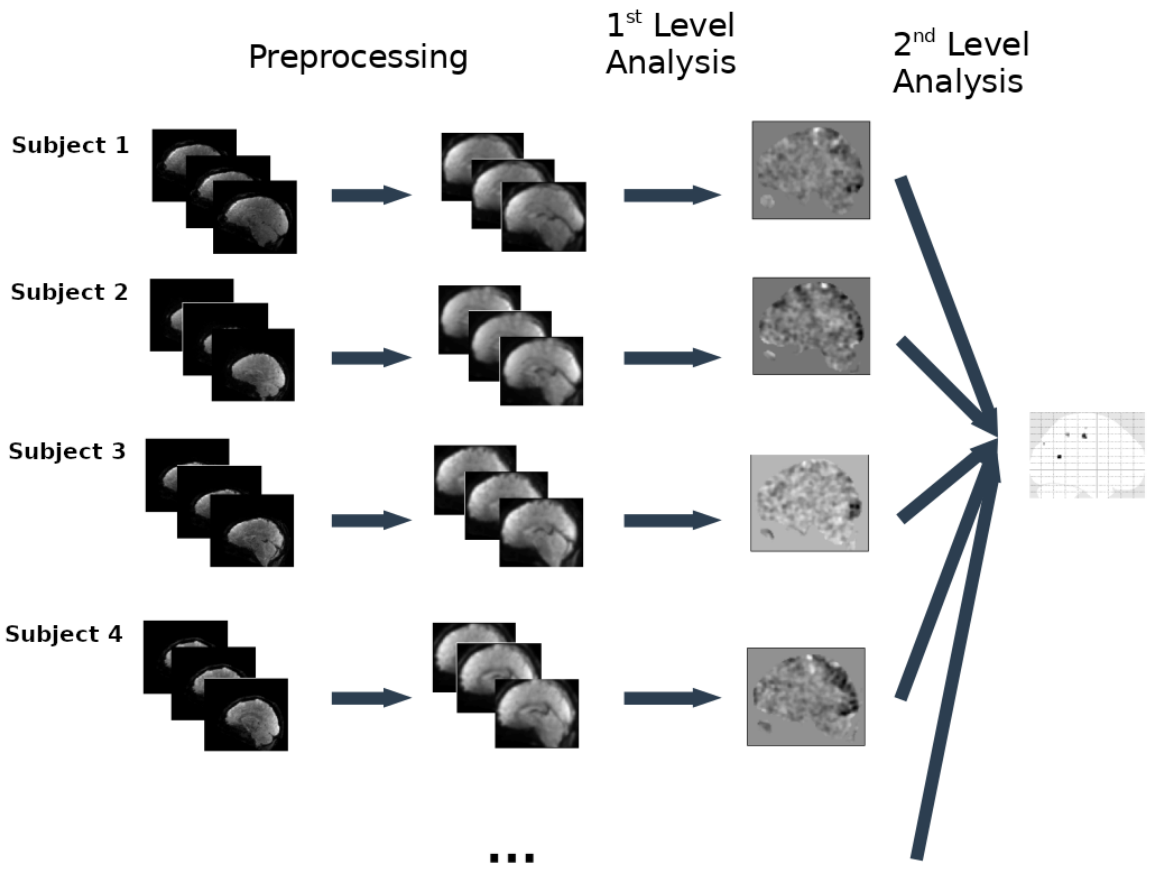


Figure 2.1 – Overview of the different steps of an fMRI analysis: preprocessing of the raw subject data, subject-level statistical analysis (first-level analysis), resulting in 3D contrast maps for each subject, and group-level (second-level) analysis with statistical inference, resulting in a global 3D thresholded statistical map.

same position in the brain, at each time point and for each subject. Since the shape of the brain varies across subjects, despite the similar structural organization, spatial transformations have to be applied on the data to fit every brain image towards a standard brain template.

For these reasons, multiple steps of preprocessing, described hereafter, are performed on the raw subject data so that subject and group-level statistical analysis can be performed. Examples of subject-level data before and after preprocessing are shown on Figure 2.2.

2.1.1 Distortion Correction

Echo-Planar Imaging (EPI) [144] is the most widely used method for fMRI acquisition. Magnetic Resonance builds images by applying a Fourier transform of a what is called a k-space. In EPI, instead of collecting one line of k-space (phase-encoding step) per repetition

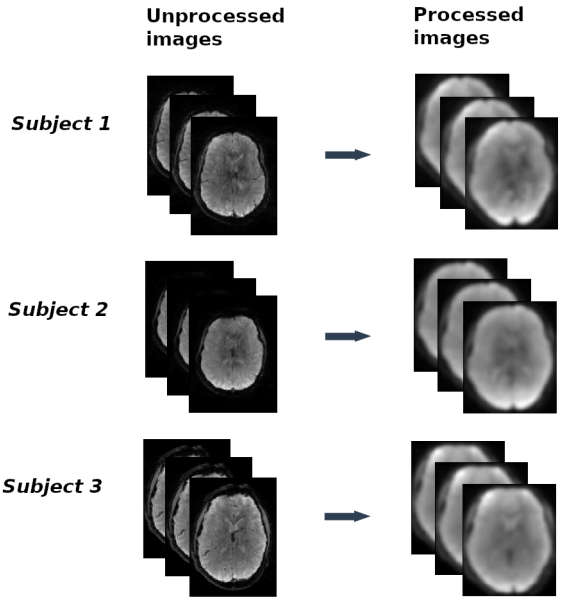


Figure 2.2 – Time-series of 3D BOLD MRI images for three participants before and after preprocessing (consisting in motion correction, registration to a template space and smoothing).

time period, as is done in standard spin echo imaging, multiple lines of k-space composing an image are collected in a single shot. This is done using two gradients, a phase-encoding gradient and a frequency-encoding gradient, which define the coordinates of the path used to fill the image. There are multiple possibilities regarding the choice of path (blipped, non-blipped, spiral for example), for which there exist multiple types of phase-encoding and frequency-encoding gradients.

One of the main problem of EPI is the presence of artifacts caused by magnetic field inhomogeneities. One of these is the dropout in signal in regions where inhomogeneity is present, which can be observed on functional images by comparing them to structural images. The other sort of artifacts caused by inhomogeneities is geometric distortion, which can be problematic in further steps of preprocessing (for example, registration, which is presented in subsection 2.1.4.

Geometric distortion, as well as signal dropout, can be limited by reducing inhomogeneities using field maps [81]. There are also methods to correct for them [76]. The main type of method for correction of geometric distortion is the use of field maps which represent the field inhomogeneities, which can be used to correct EPI images by unwarping them.

2.1.2 Motion Correction

In fMRI studies, statistical analyses are done under the assumption that a voxel's coordinates correspond to a constant position in the brain. Therefore, the time-series of intensity values

measured at given voxel coordinates is meant to correspond to the time-series of intensity values at the associated position in the brain. This assumption is used for statistical analysis, where the intensity time-series for each voxel are treated separately.

However, motion of the subject's head is almost always present during the acquisition of images in the MRI scanner. Head motion can be caused by some natural factors, such as respiration or swallowing, as well as by factors related to the study (if the study requires to perform motor tasks for example) [100]. This can induce signal variations which are caused by the shifting of the associated brain location instead of a variation of intensity at a same position. In particular, when the motion - and induced signal variations - are strongly correlated to the tasks, motion-related noise can highly interfere with the statistical analysis.

Motion correction, also known as realignment, can be done through rigid-body transformations. These are transformations which consist of translations and rotations, and which are applied to each of the 3D images in the temporal series. Given one of the images in the time-series as a target image, parameters for the translation and rotation of each 3D images are estimated by minimizing a cost function corresponding to a difference between the alignment of the brain in the target image and the transformed image. After estimation of realignment operations, the realigned image is resliced to give a value at each voxel position in the image. Multiple choices are possible in terms of target image (mean image, or image at a given time point in the sequence of images), cost function (least squares or normalized correlation ratio, for example) and reslicing (linear interpolation, or higher-order methods for interpolation such as spline interpolation) options [127]. Also, because the available and default options for these parameters vary across software packages, motion correction can be performed differently depending on the software package used [116].

2.1.3 Slice-Timing Correction

During the acquisition of MRI data, intensities at each position of the brain are not measured at the same time for a given MRI volume. Images are generally obtained using 2-dimensional acquisition, with different slices of the image being acquired at different times.

The goal of slice-timing correction is to estimate, for each voxel, its corresponding value at a given time corresponding to the time of acquisition a chosen reference slice [140]. This way, all voxels within a 3D image are temporally aligned to all correspond the same time. This is done by interpolation of the signal between two consecutive timepoints.

In the current state of research, slice-timing correction is less used because other problems can appear when applying slice-timing correction, such as the propagation of artifacts [127]. Also, problems linked to slice timing are less important with adequate slice acquisition methods, like interleaved slice acquisition, and with short times of repetition (less than 2 seconds). The impact of slice timing can also be reduced by adding temporal derivatives of the haemodynamic

response function in the general linear model at first-level analysis. Modeling considerations associated to the HRF in first-level analysis are detailed in section 2.2.3.

2.1.4 Registration to a Template Space

Motion correction, presented above performs realignment of brain images of a subject so that each voxel coordinate match a position of the brain. In order to perform group studies, voxel coordinates also need to match a same position in the brain across different subjects. Brain images from different subjects can be aligned to a common brain model, using a non-rigid transformation. This processing step is called registration (or spatial normalization) towards an atlas [127].

Registration requires choosing an atlas towards which brain images will be realigned. An atlas, built as an average of multiple brain images, gives location of specific anatomical features within a coordinate space. A template gives an image representing the atlas, towards which images can be aligned. Today, the most widespread templates for registration are the templates from the Montreal Neurological Institute (MNI templates) [17].

Registration can be carried out using the high-quality anatomical images. Anatomical images are segmented, to identify different brain tissues. Various methods exist for automated tissue segmentation, which can be divided in different families of methods, such as threshold-based, region-based, clustering or classification-based methods [33]. Registration can then be performed using non-linear transformations. A transformation towards the template is estimated from the structural image, and then used to realign structural and functional data towards the template. Figure 2.3 illustrates the steps applied for registration and the resulting structural and functional spatially normalized images. Multiple methods can be used for registration, including landmark-based methods, volume-based registration or computational anatomy [127], with default tools varying from one software package to another. It can be done at a different position in the processing pipeline depending on the software package. For example, SPM performs registration as a part of preprocessing. On the other hand, FSL does it on first-level contrast images obtained after first-level analysis.

2.1.5 Smoothing

Finally, a step of spatial smoothing is typically applied on all 3D functional images [127]. This allows to remove high-frequency noise, to increase signal-to-noise ratio, and also to reduce the remaining anatomical variations.

Spatial smoothing can be done by applying convolution using a gaussian kernel on the 3D images. Various levels of smoothing can be chosen, with the level of smoothing applied being characterized by the Full-Width at Half-Maximum (FWHM) of the kernel. The level of smoothing in the data, which depends on the level of spatial smoothing applied at this step,

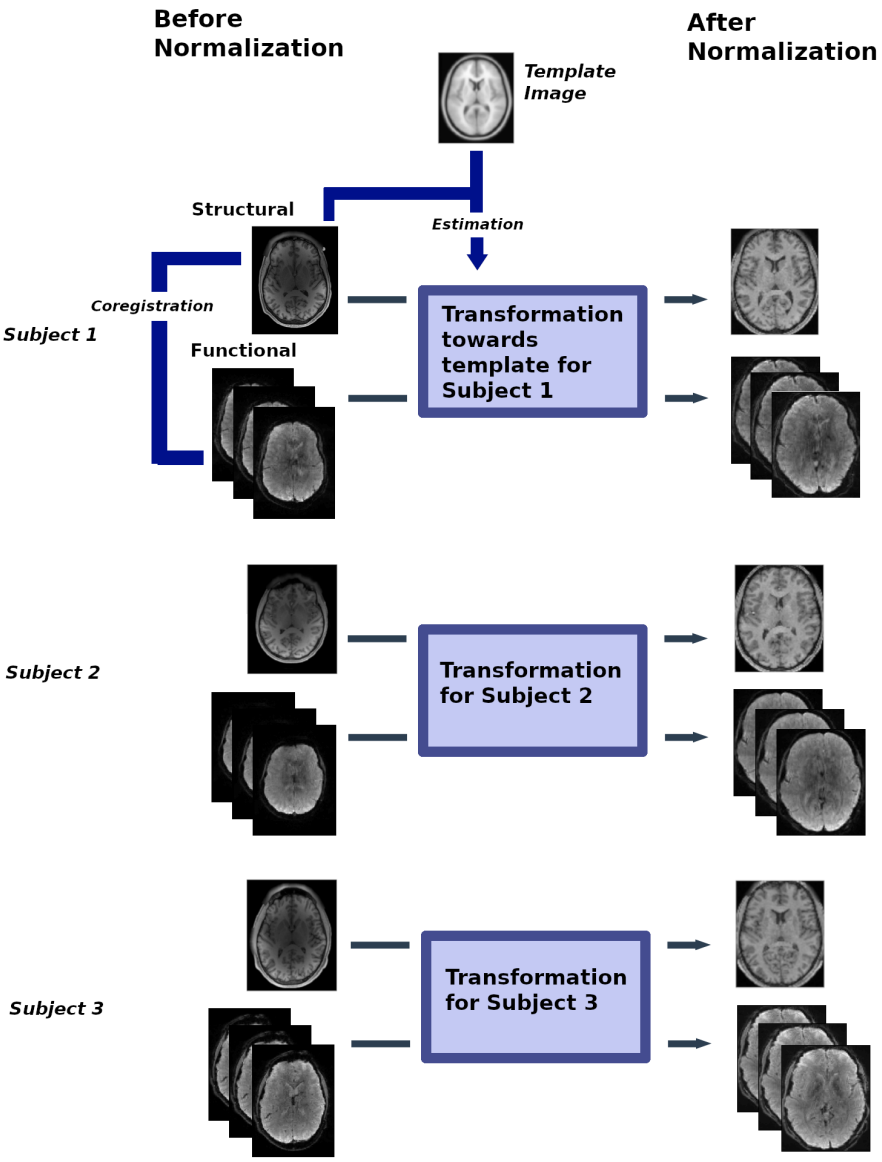


Figure 2.3 – Functional and structural images for three participants before and after registration toward the same template. Steps for registration are illustrated for the first subject data: functional data are registered and aligned to the structural image of the subject (coregistration). A transformation is then estimated using the structural image and a template image, which serves as a target for alignment of subject data. This transformation is applied on the structural image and the coregistered functional images to align them on the template.

can impact the results at statistical inference at the end of the analysis, with low levels of smoothing associated to conservativeness in results after statistical inference [72]. Examples of images obtained with different smoothing levels on a same brain image are shown on Figure 2.4.

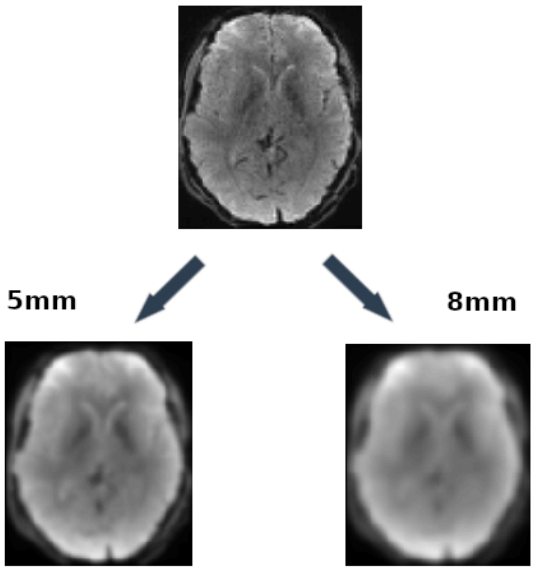


Figure 2.4 – Comparison between a non-smoothed functional subject image (top) and two derived smoothed images, with different levels of smoothing: kernel FWHM equal to 5mm (bottom left) or 8mm (bottom right).

2.2 First-Level Analysis

Once the subject data have been preprocessed, a first step of statistical analysis at subject-level, called first-level analysis, is performed. In this section, we describe the theoretical background behind first-level statistical analyses and the modeling considerations.

After preprocessing, a voxel coordinate in the image corresponds to a given position in the brain. Therefore, the time-series of intensity values for a given voxel coordinate corresponds to the variations of the fMRI signal across time at this position.

As presented in section 1.2, in BOLD fMRI, brain activity induce variations in the measured fMRI signal. Therefore, this signal can be seen as the sum of two components: a BOLD signal, which corresponds to the component of the fMRI signal whose variations are associated to the performance of tasks during acquisition, and random BOLD noise.

In fMRI, information about when subjects perform the paradigm tasks during data acquisition is given by a stimulus time-series [127]. This can be used to model expected responses in case of brain activation for each task. The goal of first-level statistical analysis is to estimate, at each position in the brain, levels of activation associated to these tasks. These levels of activation are estimated so that the sum of associated estimated responses corresponds to the best estimation of the BOLD signal component in the fMRI signal. To do this, we use a General Linear Model (GLM) to fit the fMRI time-series at each position of the brain to regressors associated to each

task.

2.2.1 General Linear Model

The General Linear Model is an algebraic tool which can be used for various types of statistical analyses. In this section, we will present the case of multiple linear regression, which is used for statistical analysis both at subject and group-level in task-based fMRI. This section presents a summary of section A.1 in Appendix A of [127], describing the theory of General Linear Model and multiple regression.

Suppose we have a vector Y of length t that we want to write as the weighted sum of $p+1$ vectors of length t (with p vectors corresponding to explanatory variables and one corresponding to constant term), and an error term. The model for multiple regression is the following:

$$Y = X\beta + \epsilon$$

with the following assumptions:

— $X \in \mathbf{R}^{t \times p+1}$ is a matrix with $p+1$ columns (regressors), with p of them representing explanatory variables and one of them a constant term. X is called the design matrix, it is non-random and known a priori.

— $\beta \in \mathbf{R}^{p+1}$ is a vector of non-random and unknown parameter values associated to each regressor, that we want to estimate.

— $\epsilon \in \mathbf{R}^t$ is a vector of random variables ϵ_i which are non-observable, constituting noise in the data.

— $Y \in \mathbf{R}^t$ is the data that we want to fit to the model, which is observable and random because of the noise added at each coordinate by ϵ.

Fitting the data to the design matrix consists in searching the estimation $\hat{\beta}$ of the set of parameter values β that best explains the data Y in function of explanatory variables X. A given parameter value β_j, can be estimated with a linear estimator $\hat{\beta}_j$, which is a linear function of the observed values Y_i. The coefficients in this function only depend on the values in X, which are known. An unbiased estimator $\hat{\beta}_j$ is an estimator which verify the condition $\mathbf{E}[\hat{\beta}_j] = \beta_j$.

Given an estimate $\hat{\beta}$ of β, we obtain an estimate $\hat{Y} = X\hat{\beta}$ of Y. The differences between Y_i and $\hat{Y}_i$ are called residuals, given by the vector $e = Y - \hat{Y}$. One possible estimator $\hat{\beta}$ for β is the ordinary least squares estimator, which minimizes the sum of squares of residuals (equal to the scalar product $e'e$). With the model here, the ordinary least squares estimator for β is given by the following equation:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

This equation assumes that $X'X$ is an invertible matrix (which requires X to have full column rank: there must not be regressors that are linear combinations of other regressors in the design matrix), otherwise there will be multiple possible values of $\hat{\beta}$ which will minimize the sum of squares of residuals.

The ordinary least squares estimator is a linear unbiased estimator of β. The Gauss-Markov Theorem states that, under certain assumptions regarding the ϵ_i, it is the best linear unbiased estimator :

Theorem 1 *If the random variables ϵ_i verify the following conditions:*

— $\mathbf{E}[\epsilon_i] = 0$ *for all* i

— *All ϵ_i have the same variance $\sigma^2 < \infty$*

— *All ϵ_i are uncorrelated: $Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$*

Then the ordinary least squares estimator is the linear unbiased estimator with the lowest variance.

In this case, an estimate of σ^2 is given by $\hat{\sigma}^2 = e'e/(n - (p + 1))$.

In task-based fMRI, for first-level analysis, multiple linear regression is used on each voxel to estimate parameter values associated to regressors corresponding to the tasks performed. Y represents the BOLD signal time-series after preprocessing and X is a design matrix containing a constant column and regressors associated to each task, built from the information regarding times of task realizations.

Other regressors may also be included in the GLM, to correct errors or capture variance in the signal, and improve the estimation of parameters associated to the original regressors.

Orthogonalization

It is common to have multiple regressors which are highly correlated to each other. When two regressors are correlated, adding or not one of them in the model can change the parameter value estimated for the other one. Also, when both are present in the model, slightly changing the signal can lead to strong variations in parameter estimates for both regressors, as some of the "explanation" for the variability of the signal shifts from one regressor to the other. Therefore, because of the variability resulting from the noise in the data, parameter estimates for these regressors will be highly unstable.

One solution to remove the correlation between regressors is to apply orthogonalization [110]. When regressors are orthogonal, the parameter value for one of them does not depend on whether the others have been included in the model. Orthogonalization consists in removing from a regressor the component which is common to other regressors.

Although orthogonalization can help avoiding the instability in parameter estimations, the interpretation of a regressor is changed when this regressor is orthogonalized, as well as the value

of its parameter estimate, since it does not represents a same portion of explained variability. Also, the shape of orthogonalized regressors – and the value of associated parameter estimates – depends on the order of orthogonalization. Depending on whether a first regressor is orthogonalized with respect to a second one or the other way around, the portion of variability which is initially common to both regressors will be attributed either to the first or to the second regressor.

Because of this, orthogonalization should be avoided as it may change the interpretation of regressors of interest, except in some specific cases. In task-based fMRI, one possible case for applying orthogonalization is for supplementary regressors which are only present to explain a component of variability and reduce error variance, such as the temporal derivatives of the HRF (which are presented in subsection 2.2.3). This is usually done in software packages such as SPM or FSL [110].

2.2.2 BOLD Signal and GLM for First-Level Analysis

In this subsection, we describe the properties of the BOLD signal which make the modeling of task regressors and the use of GLM for statistical analysis possible.

BOLD Signal Properties

As presented in 1.2, task-based fMRI uses MRI to measure a fMRI signal that varies with the blood flow associated with neuronal responses.

The expected variation of an fMRI signal caused by neuronal activity is called the BOLD signal. This BOLD signal can be described as a sum of signals associated to the tasks performed, called haemodynamic responses [97]. For a neural response to a short stimulus (as in event-related design), the temporal evolution of the response consists of the main following step [127]. First, shortly after neuronal activity (1-2 seconds), the haemodynamic response increases before reaching a peak around 4 to 6 seconds after the stimulus. It then starts decreasing until 12 to 20 seconds after stimulus, reaching a minimum slightly under baseline before slowly returning to it. For long stimuli, the signal level slightly decreases towards a plateau after reaching the peak of activation and before the undershoot. Another characteristic is an initial dip within the 2 seconds after neuronal activity, sometimes identified in fMRI studies, with low intensity, though it is often ignored in most haemodynamic response designs.

With these informations, the typical haemodynamic response can be modeled, as shown on Figure 2.5, with a function that is called the haemodynamic response function (HRF). Given an HRF and a stimulus time-series associated to a task, the time-series of the corresponding expected haemodynamic response for the task can be obtained through the convolution of the stimulus time-series with the HRF [27].

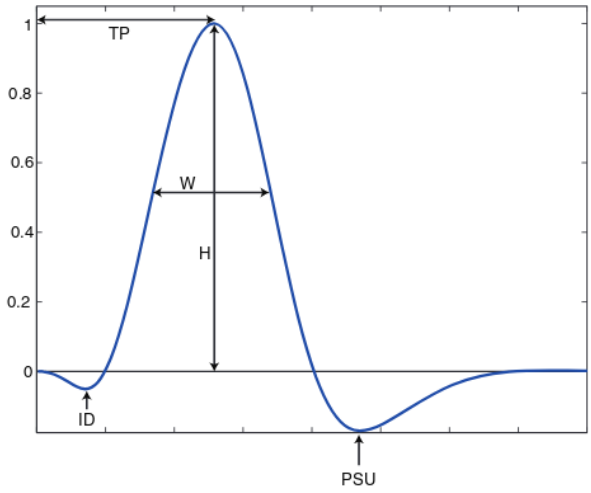


Figure 2.5 – Representation of the shape of the haemodynamic response (source: [127]). Modelling of the haemodynamic response can define the time to peak (TP) and height of response (H), width at half-maximum (W), post-stimulus undershoot (PSU) and potential presence of an initial dip (ID).

The use of convolution of the HRF for modeling relies on two properties of the haemodynamic response in function of the neural response. The first one is linearity [16, 50]: the amplitude of the haemodynamic response will be proportional to the amplitude of the neuronal response, and the expected haemodynamic response for a sum of neuronal activities will be equal to the sum of expected responses for each activity. The second one is time invariance [50]: the expected response in case of neuronal activity does not depend on the moment of activity. An example of modelling of the expected signal from the convolution of a stimuli time-series with an HRF, taken from [127] is shown on Figure 2.6.

General Linear Model for First-Level Analysis

First-level analysis uses a GLM, as described in the previous section, to estimate brain activity for each task. For each voxel coordinate, the time-series corresponding to the BOLD signal at this position is analyzed with the model $Y = X\beta + \epsilon$, where Y is the BOLD signal time-series for this voxel and X is the design matrix, containing regressors corresponding to the expected haemodynamic response associated to each task performed by the subject, as well as a constant column [53]. Figure 2.7 shows an example of design matrix obtained in SPM.

At the end of first-level analysis, a parameter value $\hat{\beta}_i$ is estimated for each regressor, corresponding to the level of activation for this regressor. A specific contrast (linear combination of parameters), corresponding to a variable of interest (for example, with the motor task, a difference of activation between right hand and left hand), can be obtained from these estimated

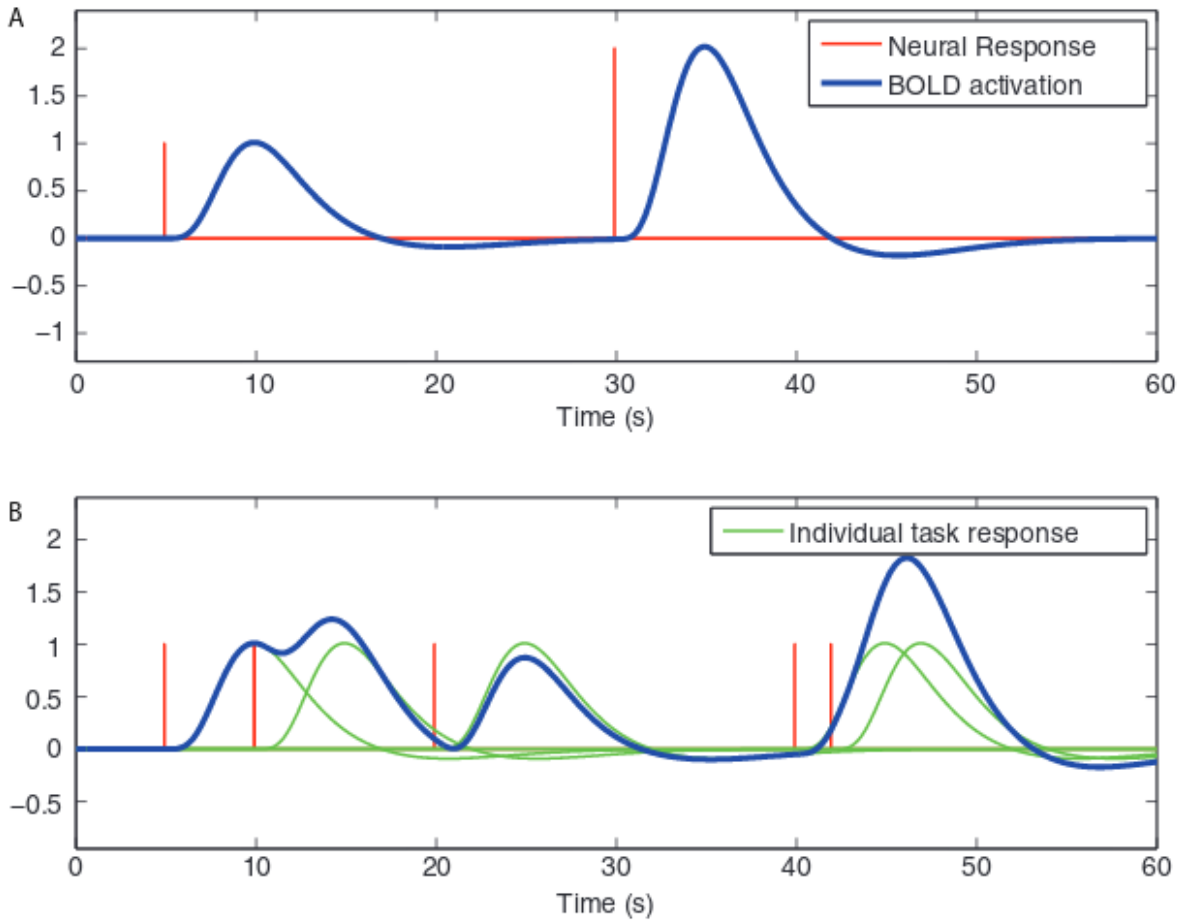


Figure 2.6 – Modelling of time-series representing the expected haemodynamic response in function of stimuli time-series in the case of event-related designs (source: [127]). Convolution of the stimulus time-series with the HRF can be done thanks to the linear time invariant property. The expected response for a stimulus (neural response in red on Panels A and B) is the same at every instant; a stimulus twice as big leads to a response twice as big (Panel A); a sum of stimuli leads to a BOLD signal equal to the sum of expected responses for each stimulus (BOLD activation in blue equal to the sum of individual task responses in green on Panel B).

parameters. For a subject, the result of first-level analysis is a single 3D statistical map for each contrast, giving the estimated contrast value at each position in the brain.

As said in 2.2.2, the regressors corresponding to the expected haemodynamic response for each task can be obtained through the convolution of the stimulus time-series with an HRF. There are multiple possibilities regarding the modeling considerations for the HRF, which are detailed in subsection 2.2.3.

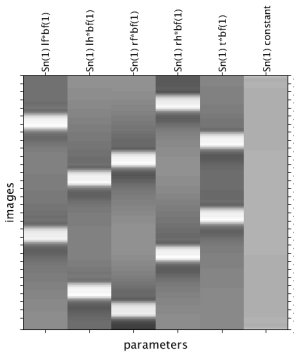


Figure 2.7 – Example of design matrix with dimensions $t \times (p+1)$, with p the number of regressors (5 here) and t the number of time-points (corresponding to the number of images in the temporal sequence of brain images), obtained in SPM for a nlock motor paradigm. The 5 first columns in the matrix correspond to the different regressors (left foot, left hand, right foot, right hand and tongue) while the last column corresponds to the constant term in the regression. The design matrix is shown as a heatmap with higher values in white. White blocks corresponding to the times of task performance can notably be seen on each regressor in the matrix.

Motion Parameters

As part of the preprocessing, motion correction estimates six translation and rotation parameters for each timepoints. Those estimates can be included in the design matrix as nuisance regressors, in order to correct any remaining motion artifacts [83]. Squares, derivatives, and squares of derivatives of the six original translation and rotation parameters can also be included, leading to up to 24 additional regressors in the model.

Researchers must be careful when using motion parameters in the model regarding the possibility that the realization of tasks is correlated with motion, in which case adding motion regressors may lead to lower parameter estimations for task regressors, as the motion regressors will "explain" part of the signal. Motion regressors can be orthogonalized to avoid this.

High-Pass Filtering

When observing BOLD signal, slow overall drift of the signal are often observed. In the frequency domain, this corresponds to a very low frequency peak. Studies about the low-frequency noise in the signal concluded that there are multiple sources for it, including effects of subject movement, or also sources which may not be attributed to the subject such as scanner effects [141].

High-pass filtering is the usual method used to correct this low-frequency noise [127]. It is performed differently depending on the software package, as various methods exist. In order to be able to apply high-pass filtering, the frequency of repetition of task blocks or events during acquisition must be higher than the highest filtered frequency, otherwise signal variations related

to BOLD activation may be erased by filtering.

One approach consists in using a set of functions that capture those low-frequency drifts within the design matrix of the GLM. The discrete cosine transform (DCT) [1] basis set can be used, where the set of function corresponds to a series of low-frequency cosine function (the highest-frequency of a function corresponding to the highest that we want to remove).

Another method consists in creating estimation of those low-frequency trends that will be removed from the data. One model which can be used to do so is the LOWESS model [26]: at each point, a linear regression is estimated locally over a window of the data, with weights on points which are higher at the center of the window (for example with a gaussian function for weighting). This linear regression is used to create values at each point of a new time-series, which can be used for high-pass filtering by capturing the low-frequency drifts.

Prewhitening

An important assumption for the application of the Gauss-Markov theorem given in 2.2.1, which states that the ordinary least squares estimator used in the GLM is the best linear unbiased estimator, is that the different noise variables at each time points are not correlated. However, even after high-pass filtering, there is still temporal autocorrelation within the data. Mathematically speaking, this can be expressed as the fact that, instead of having a noise vector ϵ such that $\epsilon \sim N(0, \sigma^2\mathbf{I})$ (with N the multivariate normal distribution and $\mathbf{I}$ the identity matrix), we have $\epsilon \sim N(0, \sigma^2\mathbf{V})$ with $\mathbf{V}$ different from $\mathbf{I}$.

One way to use the GLM under a form which verifies the assumption of uncorrelated noise is to apply prewhitening [127]. Prewhitening consists in finding a matrix $\mathbf{W}$ such that $\mathbf{WVW}' = \mathbf{I}$. Once we have found this matrix, instead of the initial problem:

$$Y = X\beta + \epsilon$$

We will solve the following problem:

$$\mathbf{W}Y = \mathbf{W}X\beta + \mathbf{W}\epsilon$$

with $\mathbf{W}\epsilon \sim N(0, \sigma^2\mathbf{I})$. Since the new noise variables are uncorrelated, the ordinary least squares estimator in this new model is the best linear unbiased estimator for β. The ordinary least squares estimator is given by the equation in 2.2.1 where X and Y are replaced by $\mathbf{W}X$ and $\mathbf{W}Y$.

The main difficulty consists in finding $\mathbf{W}$. For this, a good model for temporal correlation has to be found. The simplest known model is $AR(1)$ [162], where noise has variance equal to 1 for each time point and a correlation between two time points a units apart equal to

$cor(Y_i, Y_{i+a}) = \rho^a$ for a given value of ρ (the more distant two time points are and the lower ρ is, the less correlation there is). More models which include $AR(1)$ as a particular case also exist. Other options include unstructured estimate, which have less bias but higher variance.

2.2.3 HRF modeling

The use of the GLM requires a model of the HRF to create the regressors associated to each task. Multiple possibilities exists in terms of modeling for the HRF, with differences in assumptions and model complexity [95]. Here, we give an overview of existing models for the HRF based on the information on the haemodynamic response shape (detailed in 2.2.2).

Canonical HRF

Estimates of the HRF can be obtained by obtaining multiple responses to stimuli, and averaging them. The canonical model of HRF, which uses the distribution function of the gamma law, was estimated by [50] and [91] using this method. One common variation of this model is the double gamma function, where a second gamma distribution function is used to model the undershoot following the initial response.

Although the canonical HRF is commonly used, other modeling considerations are possible. The problem of the canonical HRF is that it does not take into account some aspects such as the variability in shape across subjects or across brain regions [70]. Because of this, it is biased towards the detection of signal which precisely fit this function only. It is possible to avoid this with more flexible models. However, having more flexible models may also result in more variance in estimates.

Finite Impulse Response Model

The goal of FIR models is to give an estimate of the HRF without assumptions regarding its shape [65]. The idea is to use a time window containing the signal, of a length corresponding to the assumed size of the HRF, with a number of time-points to estimate. Estimations of the HRF are done at each point thanks to repetitions of the signal. This allows for subject-specific modeling of the HRF. While the flexibility of the model makes it less biased, it also leads to higher variability.

Constrained Basis Sets

The use of Constrained Basis Sets consists in having a set of multiple functions that will be convolved to the stimulus onset to fit the BOLD signal, instead of just using a single HRF [161]. This solution has the advantage of being more flexible and less biased than the canonical HRF, because of the larger number of functions used to model the haemodynamic response. It also has

less variance than the FIR models, because the number of functions in constrained basis sets is usually lower than the number of time-points for estimation in FIR models. Methods to build and use constrained basis sets notably include the FMIRB (Functional Magnetic Resonance Imaging of the Brain) Linear Optimal Basis Set algorithm [161], which takes a sample of reasonable basis functions and then uses principal component analysis to build an optimal smaller set of functions from this sample.

Temporal Derivatives

One common modeling consideration in the GLM consists in adding to the design matrix temporal derivatives of the haemodynamic responses. One possible consequence of variability across subject and brain regions is that the actual time-to-peak for the haemodynamic response is shifted from the one in the canonical model. Since $X(t+\delta)$ is approximately equal to $X(t) + \delta \cdot X'(t)$ when δ is small, adding regressors corresponding to the temporal derivatives of the task regressors in the GLM can account for slight temporal differences in the time-to-peak of the haemodynamic response [70].

2.3 Second-Level Analysis

After first-level analysis, estimates of contrasts and variances of these contrasts have been obtained at each position of the brain for a set of subjects. These data can be combined for multiple subjects to perform group-level statistical analysis, which is the second level of statistical analysis in fMRI.

In this section, we will present the different models that can be used at the second level in fMRI, as well as the possible modeling considerations.

2.3.1 GLM in Second-Level Analysis

In section 2.2.2, we presented the theory behind the General Linear Model which is used for statistical analysis in fMRI.

At second-level, a second GLM is used, where the data Y corresponds to a list of contrasts obtained for a set of subjects [127]. The design matrix used to fit the series of contrasts can include regressors corresponding to a characteristic for each subject. This can be a quantitative information about the subject, like age, or a binary information where 1 or 0 is used to indicate that a subject belongs or not to a group of interest in the analysis. Similarly to first-level analysis, once the parameters are estimated, a contrast which consists in a linear combination of parameters can be obtained and used for statistical inference.

One example of a second-level analysis is when comparing effects across two groups of interest. In that case, the design matrix will contain a regressor for each group, and the value

in each regressor will be equal to 1 for participants belonging to the groups and 0 otherwise. The parameter estimate for each regressor correspond to the mean value for the corresponding group. A second-level contrast [1 -1] calculates the difference in parameter estimates between both groups. Statistical inference can be used afterwards to observe whether the between-group difference is significant in certain regions of the brain.

2.3.2 Modeling of Variance in Second-Level Analysis

Similarly to first-level analysis, estimates of variance of parameters can be obtained at second-level analysis. Notably, these variance estimates are used later for statistical inference.

At second-level, in a situation where we want to model the value of the contrast obtained within a subject as a mean contrast value within a group, plus added variance, there are multiple sources for this variance. The first source is subject-level variance, which corresponds to the contrast variance that was estimated for the subject at subject-level. The second source of variance is the between-subject variance, which may be estimated at group-level.

two main models can be used to estimate variance for group-level analysis [108]: mixed-effects and random-effects models. Mixed-effects modeling assumes that there are two components of variance responsible for the variability in the measure of a variable of interest: between-subject variance σ^2_B, and within-subject variance $\sigma^2_{W_i}$, whose effect can differ depending on subject i. The simplest model consists in assuming that the within-subject variance σ^2_W is the same for all subjects, or negligible in comparison to the between-subject variance. In this case, the mixed effect variance $\sigma^2_{MFX} = \sigma^2_B + \sigma^2_W$ is identical across subject and can simply be estimated through ordinary least squares. This method, which does not require first-level variance estimates for second-level, is notably used in SPM [49, 51].

However, in practice, within-subject variance is not always equal across subjects, and subjects with higher variance may reduce the quality of the mixed-model variance estimate. For this reason, researchers may want to consider the differences in within-subject variance across subject when modeling the variance. Subject variances σ_{W_i} for each subject i have been estimated at the first-level. They can then be used to estimate mean and variance σ_B across subjects. Estimation of mean is done using a method of weighted linear regression, called generalized least squares, instead of ordinary least squares: weights are applied for each subject data, with lower weights for subjects with higher within-subject variance [109]. Various methods exist to estimate between-group variance, including Bayesian approaches which are used in FSL [160, 7].

2.4 Statistical Inference

After second-level analysis and estimation of second-level contrast, statistical inference is applied on these data by performing hypothesis testing to detect whether or not a significant

effect of interest can be observed. In this section, we present the theory behind hypothesis testing, before giving details about how it is used in task-based fMRI studies. Notations and definitions were taken from section 7.2 in Chapter 7 and section A.2 in Appendix A of [127].

2.4.1 Hypothesis Testing

In statistics, hypothesis testing consists in considering a hypothesis H_0 about our data, called the *null hypothesis*, and determining whether or not the information present in our data give us sufficient confidence to reject this hypothesis. A null hypothesis typically consists in an absence of effect of interest. Examples of null hypothesis can be: "The mean value of a given measure is equal to μ_0 in a population", or "There is no difference in mean value for a given measure between two populations". On the opposite, an *alternative hypothesis* H_1 consists in a presence of an effect of interest.

Let us suppose that we have observations $x_1, ..., x_n$, which we consider to be realizations of random variables $X_1, ..., X_n$ (for example, values randomly sampled from a population). Variables X_i are assumed to be independent and with the same law. We want to test an hypothesis over the law of the X_i. For this, we calculate a score $t = f(x_1, ..., x_n)$ which is itself the realization of a random variable $T = f(X_1, ..., X_n)$, that we will refer to as our test statistic. There exist a a large variety of test statistics which are associated to the different hypotheses that can be tested on the data.

The goal of a test statistic is to estimate how different the observations are from what should be expected under the null hypothesis: the larger t (or $|t|$, depending on the test) is, the larger this difference is. For example, in a Student test with the following null hypothesis: "the mean value of a variable in a population is equal to a given μ_0", we expect the mean of observations $x_1, ..., x_n$ to be equal to μ_0. Therefore, with a same variance and number of subjects, the greater the difference between this mean and μ_0 is, the greater the observed value of the test statistic is.

Under the null hypothesis, the law of the X_i are known, and consequently, the law of T is known. Therefore, we can define and calculate $p = P(T > t|H_0)$ (or $P(|T| > |t||H_0)$). This value, which is called the p-value, corresponds to the probability under the null hypothesis of having a test statistic larger than actually observed. The larger t or $|t|$ is, the smaller the p-value is.

The p-value can be used to decide whether or not the null hypothesis must be rejected. A thresholding level α is defined, and when $p < \alpha$, the null hypothesis is rejected. A rejection of the null hypothesis is called a positive result whereas an absence of rejection is a negative result. In this situation, it is possible to incorrectly reject the null hypothesis when it is true and obtain a false positive result (Type I Error), or not rejecting it although it is false and obtaining a false negative result (Type II Error). With the method used to reject or not the null hypothesis, the probability of having a p-value lower than α under the null hypothesis (and thus rejecting it

incorrectly) is equal to α. Therefore, α is called the *Type I Error Level* (or simply *risk*). α is usually set at 5%.

2.4.2 Inference in fMRI studies

Voxelwise versus Clusterwise Inference

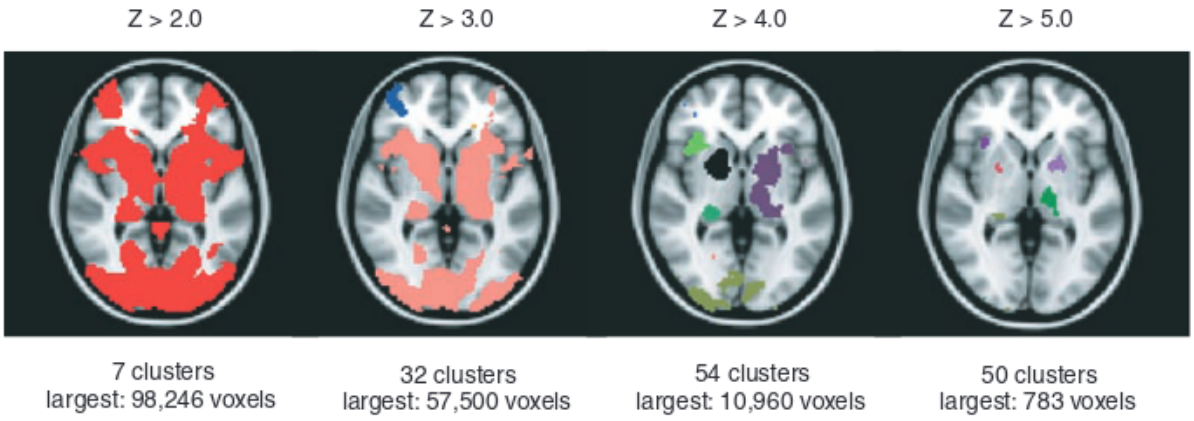


Figure 2.8 – Examples of activation maps thresholded with clusterwise inference for various levels of cluster-size thresholding, with each color representing a different detected cluster (source: [127]).

Given a contrast estimate and an estimate of its variance, a test statistic can be computed. Applying statistical inference on a second-level contrast map results in a 3D thresholded map of statistic values.

Identifying areas of significant detection can be done simply by thresholding the statistic map, with a threshold associated to the chosen risk α. There are two main possibilities regarding how thresholding can be applied on these maps [127]. The first method is to simply consider the p-value associated to the statistic at each voxel and apply thresholding on each voxel separately (Details regarding the problems of multiple comparisons and correlation in the data and their solutions are given in susbection 2.4.2). This method is called voxelwise inference.

In neuroimaging, however, areas of activation are typically smooth: statistic maps usually contain large areas of neighbouring voxels whose values are similar. Therefore, it can be more interesting for researchers to detect large connected clusters of voxels in statistic maps with levels of activation above a given threshold, instead of single voxels.

This can be done using another type of inference, called clusterwise inference, which performs two steps of thresholding to define clusters of activation. The first step consists in selecting voxels with values above a statistical threshold u_c (called cluster-forming threshold), similarly to voxelwise inference. The second step consists in selecting, among those voxels, the ones that

form sufficiently large connected clusters. This is done by defining a cluster size threshold k and selecting the connected clusters above that size. There are multiple possible definitions regarding the connectivity between voxels (sharing a face, an edge or a corner). Figure 2.8 shows an example of detections on activation maps using clusterwise inference.

Drawbacks regarding clusterwise inference include the difficulty to chose a cluster-forming threshold and a cluster size threshold [72]. Also, larger clusters can be split into multiple smaller undetectable clusters with a higher threshold, if some connecting voxels in the cluster disappear. Another problem is that when large clusters are formed, there is a lack of spatial information regarding where the activation is more specifically localized.

Correction for multiple testing

An important question regarding fMRI data is how to deal with the multiplicity of tests performed in statistical inference on the brain image. Statistic maps in fMRI can contain hundreds of thousands of voxels within the brain mask, with a statistical test applied on each of them. Because of this multiplicity of test, an analysis where the data verify the null hypothesis will almost contain activated voxels, leading to false positive results. Without correction, the expected proportion of voxels mistakenly detected as positive among negative voxels will be equal to 5%, if there is no correlation between voxels.

Because of this, statistical inference must account for the multiple comparison problem. This can be done by controlling two measures of false positives: Family-Wise Error Rate and False Discovery Rate. Family-Wise Error Rate (FWER) corresponds to the probability of having at least one false positive detection in an analysis. False Discovery Rate (FDR) corresponds to the proportion of voxels (or clusters) mistakenly detected as positives among detected voxels (or clusters). In this section, we will discuss methods used to control the FWER.

Multiple procedures can be used to control the FWER. The most simple is Bonferroni correction [127], where we use a risk $\alpha = \alpha_{FWE}/V$. With this, for V independent tests, the probability of no detection will be equal to $(1 - \alpha_{FWE}/V)^V$, which is almost equal to $1 - \alpha_{FWE}$. However, this correction is optimal when all voxel statistical values are independent, which is not the case in fMRI (because of various factors, including spatial consistency, but also image smoothness resulting from processing operations). Because of the spatial correlation, applying Bonferroni correction will result in a very conservative thresholding.

An example of method to account for the correlation within the data is the use of **Random Field Theory** (RFT), which uses topological properties within the data to define thresholding depending on data smoothness. [72] showed that using RFT can lead to conservativeness when the smoothness of the data is too low.

Another method which has been developed is the use of **permutation tests** [114]. For example, in a two-sample t-test, the statistical inference is not done only for the pair of groups

that we want to compare but also to other pairs of groups using the same subjects, where the division in two group is independent from the original division. Statistical values are obtained for all cases, and we can compare the one obtained for the initial case to the ones obtained for the permutations. If there is no group difference, the statistical value obtained for the original pair of group should not be significantly higher than for the permutations. At whole-brain level, to account for the multiple testing problem, a maximum statistic across the brain is used for the permutation test to detect whether there is any activation in the brain. This corresponds either to the highest intensity value for voxelwise inference or to the largest cluster size in clusterwise inference. Comparing these statistics allow to identify detections while controlling exactly the FWER. However, the computational cost of this solution is high, as the second-level analysis and statistical inference has to be performed multiple times.

Chapter 3

Reproducibility

Scientific progress requires to ensure that research experiments used to find new results give good approximations of scientific truth. This way, the new knowledge they bring is reliable and can be to build further research upon it [10]. One important condition for this is the ability for researchers to be able to reproduce results when repeating experiments. Unfortunately, there has been increasing concerns recently about results reproducibility in various fields of research [9, 4]. Attempts to replicate research findings gave low rate of successful reproduction [28], highlighting the weaknesses existing in the system of production and publication of research results [77].

In this chapter, we will first give an overview of issues related to reproducibility in general in scientific research (3.1). We will introduce the scientific context (3.1.1) and explain what are the issues with the current state of research regarding reproducibility (3.1.2). We will describe the various cases of attempts of experiment repetitions that may be encountered in research, the causes that have been identified in each case (3.1.3) and the solutions that have been suggested to address these issues (3.1.4). Finally, we will see how the question of reproducibility affects research in the field of neuroimaging in particular (3.2).

3.1 Reproducibility

3.1.1 Why Is Reproducibility Important in Scientific Research

The reliability of research findings is essential to scientific progress. Results from scientific experiment are used as an approximation of truth to produce new knowledge upon which further research will be built [10]. If some established results happen to be false, this will lead to question the conclusions of any research result that was based on the assumption that the previous results were true.

Experimental science works by collecting data and performing analyses, often using statistical methods. For example, in fields which involve human beings, data are acquired (through MRI in

neuroimaging for example) from a selected sample of population. Data analysis is then used to obtain results upon which researchers can make observations, possibly leading to novel findings.

Since the results from these experimental protocols are used to derive scientific conclusions, researchers may want to know whether these conclusions still hold when a given experiment is repeated. Repetition of research experiments can be done under different conditions. It may consist in trying to use the information regarding the material and methods used in a study to try and reproduce the exact same experimental results. There is also the case of result generalization, where researcher try to reproduce a result under different experimental conditions. Precise definitions for the various situations related to the repetition of experiments are detailed in 3.1.3.

3.1.2 Reproducibility Crisis

There has been growing concerns regarding the ability of researchers to reproduce research results in experimental science. This became a prominent issue in various scientific domains over the last ten years, with a number of studies showing its extent. The Reproducibility Project [28] – where multiple teams of researchers attempted to reproduce the findings of 100 psychology studies published in renowned journals – underlined the extent of this issue. Only 39 out of the 100 replication attempts were successful. Various works in other disciplines [10], such as biology [4] or drug development [9], also addressed this issue by trying to replicate findings from a sample of studies within their fields with a low rate of success.

First major concerns regarding reproducibility in modern research appeared in 2005, when [77] raised awareness towards the fact that there are many factors in experimental research frameworks that might inflate the probability that a given research finding in a published study is false and cannot be reproduced when repeating the experiment in the study. Because of this, researchers from different fields tried to replicate results from multiple studies, with low rates of successful replications. This led researchers to consider the crisis as a serious issue and question the state of research practices, as it might undermine the reliability of research results.

Attention was also given to the simpler question of the identical reproduction of research results, and whether studies gave enough information about the protocol and data so that other researchers could reproduce the exact same results. Multiple studies conducted to try and reproduce research results concluded that even this was often difficult or impossible due to the lack of reported informations [136, 10]. This induces a lack of transparency, making it impossible to detect the presence of error or scientific fraud which may be responsible for the results and undermining the reliability of these works even more.

Because of these observations, the reproducibility crisis is now seen as a critical issue in the scientific community. The fact that research findings have a high probability to be false can undermine the trust in the scientific discourse within the public opinion, which is crucial

regarding a number of scientific subjects [103].

3.1.3 Causes of the Reproducibility Crisis

The difficulty to repeat experiments and obtain similar results may be attributed to various factors which are associated to the conditions in which experiments are conducted and research results are published. An effort has been made to try and identify what may be the main reasons for this phenomenon, and what changes will be necessary in research to overcome this problem.

A first condition for reliability is to have guarantees regarding the validity of protocol and methodology applied to obtain the results. Another condition is to ensure that redoing the exact same analysis on the data gives results which are identical [136]. Doing so requires to have access to all data used in a study, as well as complete details regarding the protocol and means to reproduce it exactly. This can notably be done theoretically in cases where the analysis procedure is entirely digital (for example if it consists in the simple execution of code which is provided in a study article).

Researchers can also repeat an experiment using a procedure which is not expected to yield results identical to those of the original study. Experimental protocols are subject to multiple sources of variability including intra-subject variability (differences in data acquired at different moments), sampling of the population, method of data acquisition and analysis method. Because of these sources of variability, different experiments trying to answer a same question are not necessarily expected to find identical results. Moreover, those differences in results can lead to differences in conclusions for a same research question.

Therefore, researchers may want to know whether conclusions from an experiment are specific to the data and experimental setup of the study, or if they can be generalized under different conditions [60]. An elementary example of this is test-retest reliability, where the same whole protocol is applied on the same subject but where subject measurements may vary between both experiments, to see if there are strong variations in results. Other forms of generalization include experiments with variation with regards to the conditions of data acquisition, analysis methods used, or sampled population (or multiple variations at the same time) [113].

Because of these sources of variability, there is a variety of possible cases regarding the repetitions of experiments, each associated with different research issues. Multiple terms, such as reproducibility, replicability, or generalizability, are found in the literature to describe these various situations associated to the reproduction of experiments. However, there was a lack of clear initial definitions to distinguish the different situations they were referring to. Also, independent attempts to clarify these terms led to different conflicting definitions which can be found in the literature. Because of this, they are often used to describe a general situation, without consideration as to whether or not there are variations from the original experiment. Although there are articles that give definitions of these terms to associate them with specific

situations, there is still no global consensus regarding their definitions, and different articles may use a same term to describe two different specific situations, leading to conflicting definitions [35, 60].

To avoid ambiguity and confusion in the following, in order to distinguish the precise cases that we may encounter in terms of reproduction of experiments (and what research questions and issues are associated to them), we use the following definitions for each specific situation [158, 60, 136], which are summed up in Figure 3.1:

— *Reproducibility* of a study: repeating an experiment using the same data and the same analysis protocol, in order to obtain identical results.

— *Replicability*: repeating an experiment using the same protocol but with different data. Differences in data may be due to differences in acquisition or in sampled population.

— *Robustness*: repeating an experiment using the same data but with differences in the analysis protocol.

— *Generalizability*: repeating an experiment using both different data and a different protocol.

	Same Data	**Different Data**
Same Protocol	Reproducibility	Replicability
Different Protocol	Robustness	Generalizability

Figure 3.1 – Definitions for the possible situations of repetitions of experiments, as given in [23, 60, 158, 136]

Irreproducibility: A Lack of Data or Lack of Description

The first factors which can be mentioned are those that prevents the exact reproducibility of an experiment with the same analysis protocol and data. One of them is the absence of shared data, which is essential when trying to reproduce an experiment to the identical [128]. Although the importance of data sharing is acknowledged today in many fields of research, it was not always the case. There were no strong requirements for publication in terms of data availability, which led to low standards in terms of data sharing habits among researchers. One of the reasons for the lack of data sharing (and incentives for it) is the difficulties that can be encountered when trying to do so. Sharing data requires to have appropriate tools to make them available to the public, and for the researchers to master them. Such a context does not motivate researchers to adopt habits to manage their data in a way that would facilitate sharing them. Though data sharing is becoming increasingly important in multiple fields of research, the modification of research habits to generalize it in a convenient way is a long-term process.

Giving enough information about the analysis protocol so that the researchers may be able to

reproduce it is also a critical point. In many studies, the lack of information make it impossible to reproduce the protocol to the identical, and differences in the protocol when trying to reproduce it may cause more or less important variations on the results. Information regarding a number of more or less important steps of the protocol, from the acquisition of the data to their processing, may be omitted because they are not considered as essential. Also, even when the information is complete enough, the difficulty to reproduce an experiment may greatly vary depending on how the information is delivered (for example, whether and how code is provided to perform specific parts of an experiment) [136].

Replicability, Robustness and Generalizability

Because of the lack of reproducibility, there is a growing demand towards making the information necessary for the exact reproduction of the results available, for more transparency. Researchers could expect that, when performing experiments that can be reproduced and use valid methodology, the conclusions they infer from their results are likely to hold when the experiment is performed again under different conditions. However, even if a research finding is reproducible with the same analysis protocol and data, it does not mean that another experiment for the same research question will show similar results when using different data or analysis protocol.

Experimental sciences (including neuroimaging and task-based fMRI specifically) commonly use a statistical framework, relying on testing an hypothesis with a given risk (often 5%) of Type I error (false rejection of the null hypothesis and erroneous classification of a result as positive) [77, 136]. In this framework, if there is no effect, the probability to obtain a result that will mistakenly be classified as positive is equal to the chosen risk. If there is a real effect to detect, the probability of obtaining a result that will be correctly classified as positive will depend on multiple factors including the size of the effect and the sample sizes.

Because of variability in data (sampled population and acquisition) and analysis protocol, the results obtained - and conclusions drawn - when performing an experiment for a given research question will vary from one study to another with different configuration. Due to the conditions in which research results are obtained and published, multiple biases and practices may favor the presence of false positive results in the literature. These includes the following:

Publication bias [28, 111]: Publication bias is often identified as one of the core issues regarding reproducibility in general. Among studies under the null hypothesis, the proportion of results mistakenly detected as positives is expected to be equal to the risk value, and converges towards it when the protocol is valid. Publication bias refers to the fact that research results are more likely to be published when presenting new, positive results - including false positive results - than when presenting negative results. Because of this, the evidence for findings present

in the published literature is stronger than what should be observed in reality. This can notably be a problem for meta-analyses, where the selected results may not be representative of what real results should be.

Data Dredging and Analytic Flexibility [57, 111]: Because the statistical framework allows for the erroneous classification of negative results as positive, with a probability of 5%, testing multiple hypotheses increases the risk that at least one of them gives a false positive result. The multiplicity of hypothesis testings can result from trying to answer multiple question, or from answering a same question with multiple analysis methods. In the current state of research, performing multiple tests without accounting for the correction for multiple comparisons is less common for a single researcher or team of researchers, as it is clearly identified as a bad research practice. However, the multiplicity of analyses can still arise from the multiplicity of researchers. Notably, researchers have multiple possibilities in terms of analytic design and outcomes for a same research question. This flexibility may also increase the chance of finding a positive result by error [75]. In particular, some of these designs might be biased toward an increased chance of obtaining a false positive. The impact of this variability will depend on its extent in a field: it will naturally be lower in fields where there is a strong consensus regarding the standardization of the analysis design and therefore a smaller variety of choices.

Statistical Power [19, 77, 123]: for analyses under the null hypothesis, with a valid protocol, the probability of mistakenly classifying a result as positive is equal to the risk value, and the proportion of negatives mistakenly classified as positives converges towards this value. On the other hand, for analysis which are not under the null hypothesis, the probability of correctly identifying the result as a positive depends mainly on two things: the effect size and the sample size, which both increase the probability of detection when higher. In many fields of research, including neurosciences [123], sample sizes lower than what is considered acceptable are common, which is an important cause of the high probability that a research result is false: lower sample sizes lead to a lower probability of detection and thus a lower proportion of detected effects among analyses outside of the null hypothesis. The lower amount of detected true positive results causes a higher proportion of false results among results detected as positive.

3.1.4 Solutions

Multiple propositions have been suggested to change research practices in order to tackle these issues and improve the reproducibility, replicability, robustness and generalizability of research findings.

Publication Incentives

A first step towards more reproducibility is to have explicit incentives from journals and reviewers to reward reproducibility rather than novelty, as well as education of researchers regarding these issues and the potential sources of non-reproducibility. Concerning the exact reproduction of results, this means encouraging the systematic providing of data and protocol, presented in a way that facilitates the reproduction of results, when it is possible, for more transparency [57]. For example, code used by researchers may be provided in the study. This is, however, not always a sufficient condition: the exact reproduction of results may require to be under the same environment in terms of software packages and versions, and Operating System. Also, the exact reproduction of analysis results is not guaranteed by the providing of analysis code, in case it is poorly documented or organized for example [86].

Concerning other forms of repetitions of experiments, this can be done by promoting replication studies, especially for results of greater interest, rather than focusing on novelty, in order to counter the existing publication bias [9]. Also, more education of researchers to raise awareness regarding these issues and the potential causes of non-reproducibility and non-replicability, in their use of statistics and experimental designs, may help improving research practices.

Data Dredging and Analytical Flexibility

To overcome issues of the researcher's degrees of freedom and its consequence, such as the possibility of P-hacking, one suggested solution in various fields is the preregistration of analysis plans prior to the analysis. This practice, which consists in defining the details of the experiment that will be performed prior to the analysis, allows to ensure that researchers do not make variations in their analyses procedures which may increase their chance of finding positive results [119]. However, in practice, it may be difficult to give a clear detail of experiments before doing them [136].

Regarding analytical flexibility, attempts to define gold standards in terms of analysis is important to help reduce the variety of analysis methods, and also to avoid those which are potentially invalid. Reducing the variety of methods and converging towards standard in a field may greatly reduce the analytical flexibility for the researcher. Creating standards can be difficult though, as it requires criteria to determine what makes a method better than another. To do so, researchers can assess the validity of methods by comparing the results they give to ground truth.

Data Sharing and Statistical Power

Another important point regarding repetition of experiments is data sharing [128]: as said previously, making study data available to researchers allows them to reproduce the exact results,

which may help them being able to detect any potential fraud or error in the analysis and increase the transparency and confidence in the research results obtained [122, 151]. Researchers can also use the data that are made available to perform new, easily reproducible analyses. In various fields, initiatives for large-scale sharing of research data give researchers access to a large amount of data which can be used in new studies (for example, UK Biobank [2] provides data in various fields related to health including brain, heart and body imaging, genetics and biomarkers).

Researchers may also take advantage of data sharing to overcome the issue of low statistical power in studies within their fields. They can achieve larger sample sizes by using the large datasets available online. They may also combine data coming from multiple datasets within a same study. Researchers must ensure that doing so does not impact the validity of their studies: differences across dataset may appear, for example due to differences in acquisition sites and machines in neuroimaging.

3.2 Reproducibility in Neuroimaging

3.2.1 The Reproducibility Crisis in Neuroimaging

In many subfields of neuroimaging, including fMRI, study protocols rely on the acquisition of complex brain data, and the application of multiple steps of processing and analysis of the data in order to obtain the final results. Because of the complexity of these research frameworks in neuroimaging and the conditions in terms of research practices and incentives within the field, issues related to reproducibility, and other forms of repetition of experimental results, have been raised, with concerns about how some factors may impact the reliability of research findings.

Data Sharing

As mentioned in the previous section, the reproducibility of research results notably rely on the availability and reporting of data used in the studies. While progresses have been made in the recent years in terms of initiatives towards more data sharing [128], there is still an effort to be made in order for data sharing to become a more common and systematic research practice. There has long been several technical obstacles to data sharing in neuroimaging: first, it requires platforms which are adapted to the data. The development of projects for the storage of data [119, 63, 6], for studies at any scale in term of dataset size, is still recent in neuroimaging research. Also, it requires standards in terms of format for the information contained in neuroimaging data, including metadata for example. Researchers do not all have yet the habit of sharing their data in such formats that would make them easily reusable, notably for the reproduction of experiments [148]. This, and the fact that neuroimaging scientists come from a lot of domains and do not always all have the same technical expertise as computer scientists to share their data [128], may undermine the progress in terms of research practices for improved data sharing

compared to other fields. Finally, there are many levels of processing to be applied on the data, and thus many levels of derived data which may be shared, including preprocessed data and first-level statistical maps at subject-level, and second-level statistical maps and activation maps at group-level. Raw subject data are the most reusable data, but it is also more costly in terms of storage to share them than just sharing the activation maps, which are the final derived data resulting from the analysis [125].

Code Sharing

Another issue regarding the reproducibility of research finding is the ability to repeat the exact protocol applied on the data. A major problem regarding methods reporting is that the protocol for neuroimaging studies requires multiple steps for which there are multiple possible options. A lack of information regarding the exact details of the protocol will prevent the reproduction of the analysis, as it will be very unlikely that an attempt to redo the experiment will not differ from the original on some points. [21] showed that, in a sample of 241 fMRI studies, reporting of detailed information regarding certain crucial steps in the experimental design, data acquisition, processing and analysis of the data was missing in an important proportion of studies, making their exact reproduction impossible. Because of the lack of reproducibility, as in many other fields, there has been more demand for and incentives towards the release of code used to perform analysis in neuroimaging studies. Also, fMRI studies are performed using software packages whose versions may change, and software version can lead to changes in algorithm used to perform certain operations [15]. Similarly, Operating Systems and their versions may perform calculations in different ways [58], which may lead to important changes when applied on successive operations.

Analytical Flexibility

Analytical flexibility has been mentioned in the previous section as a key problem regarding the reliability of experimental results. This problem is notably important in neuroimaging, and functional MRI in particular. Pipelines are composed of an important number of processing steps for which various methodological choices are possible. These methodological choices notably include choices regarding performing or not certain steps (in preprocessing, for example), choices in parameters or algorithm used to perform these steps, and choices regarding the order in which they are performed (which can be linked to the way a software package or another will perform the processing steps). This high number of possibilities at each step leads to an important number of possible pipelines. Because of this, multiple researchers or teams of researchers will very likely use different pipelines when performing an analysis for a similar paradigm independently: in [14], where multiple teams of researchers were asked to perform an analysis for a similar paradigm, pipelines were all different across teams. [20] found that, using 6912 pipelines on the same

dataset, under the null hypothesis, approximately 90% of voxels showed activation for at least one pipeline when using uncorrected, showing the strong impact of analytical flexibility on the possibility of finding false positive results in neuroimaging studies.

Statistical Power

Finally, statistical power has been identified as one of the main causes for the increased rate of false positive results in neuroimaging studies. Lower sample sizes and statistical power among studies implies a lower probability of detection of true positive results and an increased probability that a given positive finding in the literature is false [19]. Since the probability of detecting effects depend on effect size and sample size, the smaller an effect is, the more critical sample size will be for detecting it. [123] shows the evolution of average sample sizes in neuroimaging studies until 2015, and points out that statistical power is still too low to find reasonable effect sizes: in 2015, the median sample size in single-group fMRI studies was estimated at 28.5, corresponding to a median effect size associated to 80% statistical power equal to 0.75. While there has been an increase in sample sizes in recent years, there are still progresses to be done.

3.2.2 Solutions in Neuroimaging

Publication Incentives

Concerns regarding reproducibility in neuroimaging have led to more requirements in terms of transparency from researchers in the community. One key element for transparency is the release of data and code in order to allow for the exact reproduction of the experiment [128, 119, 63]. Because of the way it has been developed, neuroimaging is a field where all processing of information after the acquisition of data is digital. Due to the complexity of neuroimaging pipelines, and the important number of variable parameters, the only way to provide full information regarding the analysis protocol is to provide the code used for it. It is not sufficient to be able to reproduce the protocol completely, however, as their execution may require a specific work environment and software and operating system conditions. Incentives for practices to improve reproducibility of experiments in neuroimaging may come from exigence from funding agencies or journals. Individual motivation from researchers for better research practices may also come from the fact that it may lead to higher confidence from the research community: it has been shown that studies with shared data are prone to have less statistical errors and higher effect sizes [159]. Besides code and data sharing, it is necessary that researchers give a clear and exhaustive reporting of information regarding the experiment when describing it in their articles, so that other researchers may have a full comprehension and insight of the analysis performed [113].

Data Availability

Neuroimaging has been one of the fields doing the most efforts to improve data sharing practices [154]. An early example of data sharing initiative in task-based fMRI is the Functional Magnetic Resonance Imaging Data Center (fMRIDC) [155, 153], established in 1999, to allow sharing of functional neuroimaging data. This project was not well received at first, with researchers reluctant to share their data for various reasons, such as data protection. Initiatives for providing neuroimaging data nowadays include projects such as OpenNeuro [124], XNAT Central [74], 1000 Functional Connectomes [12, 104], Human Connectome Project (HCP) [152], UK Biobank [2] or ADNI [107]. Some of these projects give access to large sets neuroimaging data collected for specific studies made available publicly. Others are databases which can be used by researchers to make their data available online.

For task-based fMRI, as well as for other neuroimaging modalities, the multiple steps of processing (2) leads to multiple levels of data which may be shared: raw data, or derived data which include processed subject data, statistical maps, and coordinate-based data showing the zones of detected activation [125]. Given that the code used to apply processing steps on it is provided, and the environment on which it was executed can be accessed with its specific software and operating system conditions, raw data can be used to obtain all derived data. On the other hand, derived data is more closely related to the results, and providing it avoids other researchers to perform the whole processing just to have it (besides the question of the simple reproduction of research results, it also has advantages such as its interest for data re-use). Some data sharing platforms are used for specific types of data (NeuroVault [64] for statistical maps for example).

Finally, efficient data sharing requires good organization of data and meta-data. The main standard of organization is the Brain Imaging Data Structure (BIDS) [62], inspired by the data organization in OpenfMRI. Such data organization is necessary to enable data re-use, either for the reproduction of an experiment or for other reasons.

Protocol and Code Sharing

Besides data, the second requirement for the reproduction of experiments is making the exact protocol applied on the data available to the researchers. Because of the fact that the processing performed on acquired raw data is digital in neuroimaging studies, availability of protocol essentially means the availability of code associated to the operations performed on the data. All information related to code can be made available being released online. One way of doing so is by storing it in online repositories dedicated to this, such as Github. One issue with code sharing is long-term accessibility. In particular, only sharing URLs may not be sufficient to ensure that code is accessible, as they may expire [113]. Platforms such as Zenodo have been developed to allow for long-term access towards digital material, including code, using Digital

Object Identifiers. However, same as for data, sharing code is not a sufficient condition in itself to make it re-usable [39]. Good documentation and quality of code are necessary for researchers to understand it. Neuroimaging software packages can save pipelines in the form of batch or scripts, which can be re-used for exact reproduction of processing steps. This allows to have any information that may be missing when trying to describe the whole protocol details in methods reporting sections of research articles. It is also essential to make the reproduction of experiments easier using automation. Frameworks such as Nipype [61] can be used to try and replace manual interventions in the reproduction of the protocol when it is possible, by transforming it into a workflow. These frameworks can be combined with tools for "literate programming" like Jupyter Notebook [88], where researchers can use an interface combining the detailed presentation of the analysis with directly executable code, to give a better understanding of the role of each step within the workflow. Finally, the exact reproduction of experiments through execution of code requires access to the environment in which the code was executed when it is possible. Virtual machines or equivalent may be used to allow the reproduction of results on a consistent environment. Initiatives such as NeuroDebian [68] aim at providing such material to improve reproducibility.

Material and Methods Reporting

Finally, while data and code sharing is essential to ensure the reproducibility of neuroimaging research results, it is meant to be associated with a clear reporting of experimental details within research articles. The Committee on Best Practices in Data Analysis and Sharing (COBIDAS) [113], which has been developed by the Organization for Human Brain Mapping (OHBM) to address concerns regarding reproducibility, has given guidelines in order to make more reproducible research. Besides practices regarding sharing of code and data, a focus was made on the necessity to give all details regarding each part of the experiment: subject selection, data acquisition, processing applied. In particular, full detail regarding the analysis and what statistical model was used is necessary for the reader to have a full insight and comprehension of the study methodology. Also, details regarding software and operating system version used are necessary to make it possible to reproduce the experiment exactly, as algorithms and calculations may differ depending on the version.

Analysis Methods

Preexisting or new analysis methods have to be tested on real data in order to have some guarantees regarding their validity. Datasets may be used to try and apply new methods on them [125]. Re-using data from available datasets may allow to do these types of studies more easily. Specific frameworks for the optimization of pipelines, notably by estimating performance metrics associated to various factors including reproducibility [139, 89, 145]. If it is possible,

having converging optimized standards in terms of operations performed in the analysis will lead to the exclusion of certain choices of operation, algorithms or parameter values, and thus reduce analytical flexibility and researcher's degrees of freedom. More specifically, it may result in the exclusion of combination of processing choices which may be responsible for a higher risk of false positive results. Therefore, this may reduce the extent of issues related to analytical flexibility.

Data Re-Use and Statistical Power

Data sharing has been pointed out as a necessity for researchers to make their studies reproducible. Another advantage of data sharing is the possibility of re-using data from previous studies in new studies. This may concern databases collecting datasets from multiple studies, as well as from large datasets such as the Human Connectome Project. Pipeline evaluation, as described previously, is one example of possible case for data re-use to address issues related to the reproducibility crisis. Besides questions linked to reproducibility and other forms of repetitions of experiments, data sharing is known to have multiple advantages: it gives greater potential of use for a dataset, as it allows researchers to use it within frameworks they were not originally meant for; it leads to economies in terms of data acquisition; and it allows researchers to ask questions which were originally impossible to answer [128, 125]. Finally, the problem of low statistical power, which has been described as one of the main problems in neuroimaging regarding the reliability of research results, can be tackled by taking advantage of the large datasets available to increase sample sizes in fMRI studies [125]. Furthermore, the possibility to combine data from various datasets would allow to have even bigger sample sizes.

3.3 Conclusion

In this section, we have given an overview of the state of research regarding reproducibility and other associated concepts such as replicability, robustness or generalizability, presented what are the main identified causes of the reproducibility crisis and what solutions have been suggested.

Causes for the lack of reproducibility are the absence of data sharing and good protocol sharing. Stronger requirements from journal and initiatives to make these easier for researchers may partly solve this problem in the future. Causes for the lack of repeatability of results under different conditions include the bias in publication towards novel positive results rather than attempts to replicate or generalize results, the variability in analysis protocols which allows a variety of analyses and inflates the probability of finding false positive results, as well as the low statistical power in many fields of studies. Ways to solve these issues include changing publication incentives in journals, optimizing and standardizing analysis methods and achieving

larger sample sizes, for example with data re-use.

We have seen how neuroimaging is concerned by all these issues and suggested solutions, and what initiatives have been taken in order to address these issues. In the next chapter, we will focus on the issue of analytical variability and its impact on neuroimaging studies.

Part II

Contribution

Chapter 4

Data Compatibility

In the first part, we introduced task-based fMRI, how it can be used to study brain function, and we gave a detailed description of the experimental protocol used in fMRI studies, and the processing and analysis steps applied on the data in order to obtain scientific findings. Finally, we also discussed an important issue in scientific research: the reproducibility crisis.

One of the causes of the lack of reproducibility in neuroimaging is the lack of statistical power induced by low sample sizes. In recent years, various data sharing efforts have been initiated in neuroimaging, leading to a large amount of data available for re-use. In order to improve reproducibility in neuroimaging studies, we suggest to take advantage of the large number of shared datasets and re-use existing data to increase sample sizes. However, this can lead to a situation where researchers have to combine data processed differently at the subject-level.

In this chapter, we describe the present situation regarding statistical power in neuroimaging studies and we propose to improve on this by reusing and combining shared datasets. We then discuss the issue of analytical variability in neuroimaging, and describe open research questions regarding its impact on neuroimaging results.

4.1 Sample Sizes and Statistical Power

Task-based fMRI group studies rely on statistical analyses using subject data to perform statistical inference at each position of the brain (cf. chapter 1 for more details). In hypothesis testing (cf section 2.4.1), there is a possibility of mistakenly rejecting the null hypothesis, i.e. detecting effects when there is none (false positive results) or, on the opposite, not detecting an effect when there is one (false negative results). Statistical power can be defined as the probability of correctly rejecting the null hypothesis, i.e. detecting an effect in a study when there is one [127]. When performing hypothesis testing, the probability of falsely detecting an effect when there is none is by construction equal to the chosen alpha level (typically $\alpha = 0.05$).

On the other hand, when there is an existing effect, the probability of detecting or not this effect depends on multiple factors. This includes effect size (the stronger the effect, the easier it is to detect) as well as sample sizes (higher probability of detection with high sample sizes).

When performing a study with a given sample size, statistical power may be computed for a given effect size. In the simple case of a single hypothesis test, statistical power is easy to compute; however, in neuroimaging studies, researchers have to account for the multiple comparison problem and correlation in the data. Computation of statistical power in fMRI is still an open field, with multiple methods for the various possible types of studies that exist in fMRI [37].

Low statistical power has been identified as one of the leading causes of non-reproducibility in experimental research, including neuroimaging [77, 19]. Various reproducibility issues associated to low sample sizes and low statistical power have been discussed in the literature. The main problem of low statistical power is its impact on the proportion of true positives among results detected as positives. With low statistical power, the proportion of detections in studies with existing effects is low. A lower number of detected real effects lead to a higher proportion of false positive results among results detected as positive [132]. Also, various factors may lead to an overestimation of effect sizes [157]. This includes publication bias: results have a higher chance of being published when they are positive and new [111]. Another one is selective reporting: researchers may choose to not report information that is not deemed to be relevant about the results, although it may lead to a bias that changes the interpretation regarding the outcomes of interest [111]. Beyond low statistical power, other statistical issues arise due to low sample sizes. For example, [156] showed that for cross-validation, large error bars where found in fMRI studies, with an association between larger error bars and lower sample sizes.

[147] observed the median effect size found in neuroimaging studies, and for effect sizes of various amplitudes (small, medium or large compared to the found median effect size), estimated the statistical power of psychology and cognitive neuroscience studies. They found that the statistical power was low for medium effect size (0.44), and also that cognitive neuroscience studies had lower statistical power than psychology studies, suggesting that attempts to reproduce studies in cognitive neuroscience as has been done for psychology may give low rates of successful reproduction. [123] observed the evolution of sample sizes in neuroimaging studies (Fig. 4.1), finding an increase in median sample sizes, as well as an increasing number of studies with high sample sizes (more than 100 subjects). While this suggests that there is an evolution towards better research practices in terms of reproducibility, improvements are required to achieve larger sample sizes in the future.

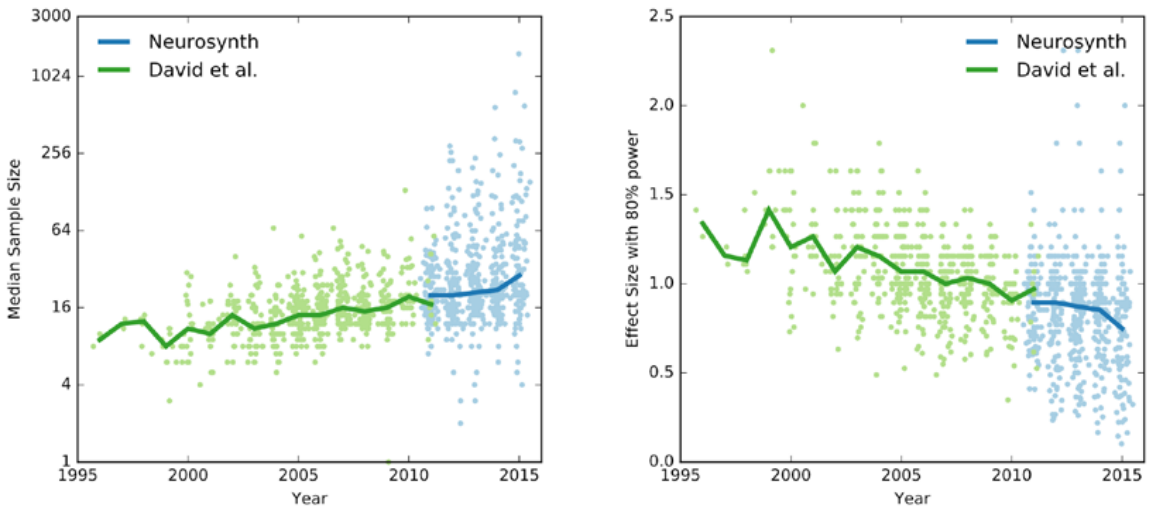


Figure 4.1 – Evolution of the median sample size (left) and median effect size for 80% statistical power (right) in fMRI studies, with sample sizes shown on a logarithmic scale (source: [123]). Sample sizes were extracted from meta-analyses by [31] and from the NeuroSynth database.

4.2 Increasing Statistical Power: Combining Shared Data

Another evolution in research practices in recent years is the increase of available, re-usable shared datasets. We mentioned in section 3.2.2 a number of data sharing projects in neuroimaging, which give access to a large amount of data for researchers to perform new studies. In this section, we describe the development of data sharing initiatives and the current state of the art in neuroimaging, as well as the possibilities in terms of data re-use to provide solutions for the lack of statistical power.

4.2.1 Data Sharing in Neuroimaging: History and State of the Art

Data sharing emerged in the field of neuroimaging in the 2000s with the fMRIDC [155, 153]. Before that, data sharing was not encouraged in neuroimaging research, for various reasons, including protection of data or difficulties in data preparation and organization for data sharing. It was the emergence of important research results made possible with data sharing that made the research community aware of the potential of data sharing and motivated the development of new projects.

Since the 2010s, an important number of new data sharing projects have been developed in neuroimaging, along with data sharing practices to make them efficiently re-usable. Shared datasets are acquired for different purposes and may include different types of data. [90] gives an overview of the main current data sharing projects in neuroimaging today, divided in three

categories.

The first type of data sharing projects is repositories to share previously existing data. Examples include the 1000 Functional Connectome Project, now International Data Sharing Initiative (INDI) [12, 104], for sharing resting-state data. Similarly, the OpenfMRI Project (now OpenNeuro Project) [102], which started in 2013, has been used to collect data from neuroimaging datasets, mainly in task-based fMRI initially, but also for other neuroimaging modalities nowadays.

Other types of data sharing efforts include projects where large sets of new data are acquired and shared. These notably include the Alzheimer's Disease Neuroimaging Initative (ADNI) [107], which started in 2004 and collects data for subjects with Alzheimer's disease, for a multisite longitudinal study; the Human Connectome Project (HCP) [152], which collected data from 1,200 healthy young adults, and influenced the development of other large-scale projects; and also the UK Biobank [2], which contains biomedical data for a longitudinal study on more than 500,000 adult subjects.

Finally, other data sharing projects include datasets collected locally, usually with a smaller number of participants, but for which a large array of data can be collected. These datasets are called densely sampled datasets, and can notably be used to study within-subject variability in detail. The number of such datasets is increasing rapidly. The MyConnectome dataset [126], which contains 104 imaging sessions over 18 months for a single subject, or the StudyForrest project [71], which includes structural data and functional data for various tasks for 36 participants, are examples of densely sampled datasets.

These different types of data sharing initiatives give researchers access to a large amount of data, in terms of both numbers of subjects and amount of data per subject. Also, besides these initiatives, there is still also an important number of fMRI studies published each year which could increase the amount of available shared data if data sharing became a more generalized practice. This suggests that an increasing number of research questions could be addressed by re-using existing data in the future.

4.2.2 Re-use of Shared Data and Combination of Data

Multiple situations may appear where re-using and combining data can be an appropriate way of performing studies with large sample sizes. This notably include meta-analyses for which result data from multiple studies are used in a same analysis. The discovery of similarities between networks derived from Independent Component Analysis in resting-state and task fMRI data [143], using activation maps from task data in the BrainMap database, is an example of research result made possible by data re-use.

A possible situation of data re-use is using multiple datasets from different studies which can be combined within a same study to address a new research question (for example, differ-

ences in anatomical images between healthy subjects and participants suffering from neurological pathologies). The median sample size in neuroimaging data is still slowly increasing. The possibility of combining data from various small datasets could allow a substantial increase in statistical power.

Very large-scale data sharing projects exist, such as UK Biobank, suggesting that large sample sizes could be achieved with data from a single shared dataset instead of combining different ones. However, due to the large amount of possible research questions that may be addressed by researchers, there can be situations for which even these very large datasets cannot provide a sufficient amount data for these new studies, or even any data at all. This can be the case when addressing specific cognitive processes which are not covered by the tasks for participants in these projects, for example. In these situations, combination of data may be the only way of achieving large sample sizes through the re-use of shared data.

Finally, data acquired for a specific purpose can also be used to address other research questions. Data sharing can be used to exploit the full potential of existing data, and doing so when it is possible may become more important for sustainability. Reusing shared data would avoid having to acquire and process new data for a single use case, as was commonly done in research before data sharing initiatives became more popular, and is still a widespread approach today.

4.3 A New Challenge: How To Assess The Compatibility of Data Processed Differently?

4.3.1 Data Combination and Subject-Level Processing

While the combination of subject data coming from different datasets may be an interesting possibility in order to increase sample sizes and statistical power in fMRI studies, this requires to first assess the compatibility of the datasets available. We detailed in chapter 5 multiple sources of variability in fMRI studies, such as population or technical variability. Some of them may occur at subject-level and create differences across subject data when combined from various datasets. Some sources of variability, such as technical variability, have been studied, and their impact on the results is known, with methods developed to correct it [105]. However, this is not the case for analytical variability, that we study here. In our work, we addressed the question of how analytical variability could impact the results in this situation.

In fMRI, the standard way of performing group studies is to use contrast maps from multiple subjects obtained after first-level statistical analysis on the preprocessed subject data. All subjects' raw data are supposed to have the same preprocessing and first-level analysis applied on them. However, when re-using shared data, available data for each subject may only include data derived up to a certain point (for example, only contrast maps resulting from first-level

analysis), instead of raw data, as there are many levels of processing applied on subject data. When this is the case, data combination may require to use, for group studies, subject data which have been processed using different subject-level pipelines, as shown on Figure 4.2.

In a typical neuroimaging study, identical subject-level preprocessing and analysis is applied on all subject data. However, there are multiple possible reasons for the unavailability of raw data, which would prevent us to do so. One of them is sustainability: raw data and derived data constitute a heavy set of data for each subject. As storing online the exhaustive digital content related to each subject may be costly, researchers may choose to only store a limited amount of data. For example, they may only share statistical maps at the end of subject or group-level analysis and not raw or intermediate processed subject data. Another major reason is privacy: legal requirements for data sharing can include conditions regarding how subjects can be identified with the data shared, as it is the case in Europe with the General Data Protection Regulation. To reduce the possibility of identification, researchers may choose to only share data derived from the raw data up to a certain point.

4.3.2 Research Question: What Is The Impact of Analytical Variability On Data Compatibility?

In this thesis, the main question that we want to answer is whether or not differences across subject-level pipelines may cause invalidity of the results obtained when combining those data in a group-level analysis. We may also want to determine, when differences across pipelines have an effect on the results, what are the specific differences between pipelines which are responsible for the effect on results, and the effects of combination of parameter differences. Finally, once we have assessed the effect of pipeline differences, we may want to use methods to correct them and evaluate them.

To answer these questions, we defined a framework to evaluate the validity of analyses combining subject-data processed using different pipelines, first without applying correction methods and then with correction. Chapter 5 presents the methodology that we use to address the research questions and how our work differs from existing works regarding analytical variability in neuroimaging. Chapter 6 presents our first contribution, in which we assessed the effect of various sources of variability on the validity of between-group analyses combining data processed differently. Finally 7 presents the results obtained when applying a method to correct the effect of analytical variability in analyses similar to ones performed previously.

The variability associated to the analysis procedure applied on the data to obtain results is called analytical variability. In order to contextualize our work, in the next chapter, we will describe analytical variability in details, how its impact on result has been studied in various frameworks and how our research framework differ from the literature.

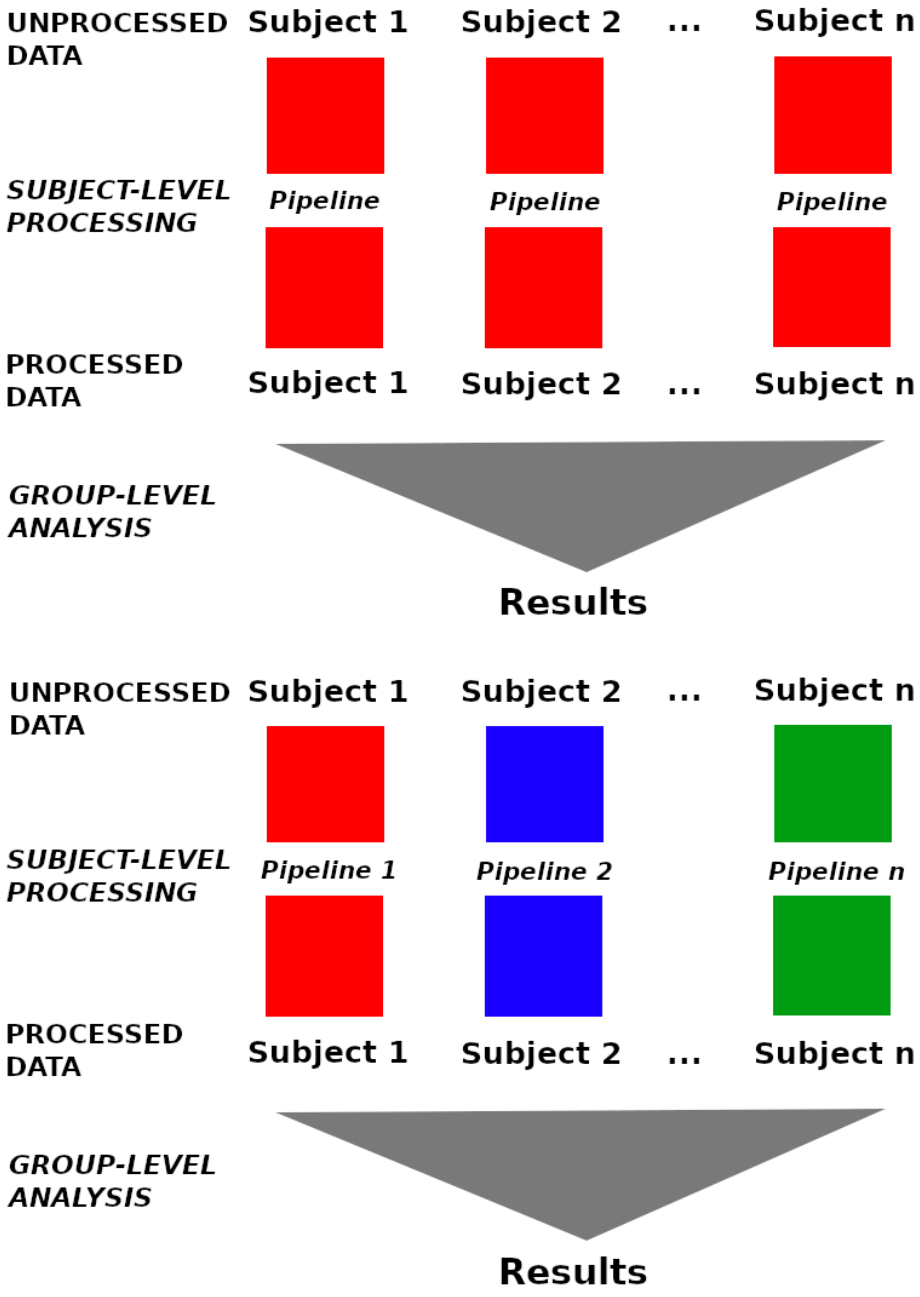


Figure 4.2 – Example of study involving combination of data generated with different pipelines (bottom), compared to what is usually done in fMRI studies (top): group-level analysis is performed between subject data already processed with different subject-level pipelines. These differences in subject-level processing may lead to differences in results compared to what we would have if all subjects had been preprocessed the same way, as is usually done.

Chapter 5

Analytical Variability

In chapter 3, we discussed the issues related to the reproducibility of experiments in neuroimaging. We saw that low statistical power induced by small sample sizes was one of the main reason for the lack of replicability of experiments. We then suggested in chapter 4 that this issue could be tackled by reusing existing shared datasets, in order to achieve bigger sample sizes.

However, re-using existing datasets may imply combining data which have already been processed. If differences in analysis protocol applied on the data impact the results, this may reduce the interest of the increased statistical power. The variability induced by different protocols and methods applied on the data is called analytical variability. We saw in chapter 2 that in fMRI, there are multiple possible choices for each processing step within a pipeline, leading to an important number of possible pipelines available to answer a given research question.

The impact of the different types of variability on research results has been studied in multiple contexts. In this chapter, we will detail the potential types of variability existing in neuroimaging studies. We will then focus on analytical variability, its potential sources in terms of methodological choices within pipelines and software conditions and its impact on the results. Finally, we will present a set of open research questions related to analytical variability.

5.1 Types of Variability

The study of variability, how it affects results and how to potentially take its effect into account is a well-covered topic in experimental sciences. In this section, we will define the main sources of variability which can be observed in general.

5.1.1 Inter-Subject Variability

Inter-subject variability is the most studied source of variability that researchers may want to address in neuroimaging studies. Different participants may have important differences between

them regarding certain features [59]. This includes variations in brain morphology, which can be addressed with spatial registration, but also in functional organization [130, 149]. In fMRI, the question regarding inter-subject variability is addressed during second-level analysis, with mixed-effect modeling.

5.1.2 Test-Retest Variability

We defined test-retest variability in chapter 3 as the variability observed on results over repeated measurements for a same subject, with same instrument and site for data acquisition and same complete protocol applied on the data. Test-retest reliability is important to ensure that a given result can be reproduced with the same subjects and protocol. An important within-subject variability could lead to question the results obtained in a study as it would suggest that important changes in results could be observed when doing an experiment multiple times.

Various potential situation appear where researchers want to measure test-retest variability. It is notably important when trying to identify biomarkers for diseases [41, 101]. Another situation where test-retest reliability can be considered is the observation of variations in measurements over a long period of time [3].

5.1.3 Inter-Scanner and Inter-Site Variability

In neuroimaging studies in general, inter-scanner and inter-site variability corresponds to a general problem in research, called technical variability [47]. Multicenter fMRI studies have the advantage of allowing to reach larger amounts of subjects in situations where it is difficult to find sufficient participants in a same place (for rare pathologies for example). However, doing so results in adding sources of undesired variability, such as scanner effects, which constitutes a drawback. The problem of scanner effects and inter-scanner variability also exists and have been studied for other neuroimaging acquisition modalities, such as PET scans or Computed Tomography (CT) scans [99].

There are different possibilities in practice regarding how researchers can account for this variability. They may either consider it as random variance, or use methods to reduce it [48]. Methods to estimate and reduce technical variability have notably been developed in genomics [82, 55], and have been adapted to fMRI [47]. These correction methods can notably be used in multicenter studies for the assessment of inter-scanner differences and scanner effects [48]. This can be done by having data for various subjects obtained on multiple MRI scanner (accounting for within-subject variance) [11].

5.1.4 Analytical Variability

Finally, the source of variability which will interest us in the rest of the chapter is analytical variability, which consists in all variability related to the operations applied on the data after the acquisition. We have seen in the first chapter that neuroimaging pipelines, in task-based fMRI, consists in a number of processing and analysis steps for which multiple options are possible. This multiplicity of options leads to a large amount of possible pipelines, which may give different results for a same analysis framework. Moreover, different software packages and operating system conditions can also have an effect on the results obtained.

Like other sources of variability, analytical variability is a common issue in a number of fields, and in various neuroimaging modalities. In MRI, issues related to pipeline variability have been shown for cortical thickness measurements [66]. Difficulties exist in the assessment and correction of variability related to the analysis procedure, compared to variability related to subjects, population or scanner effects. As shown in [115] for PET, in particular, it is difficult to identify the effect that each methodological choice has on the results.

In the following sections, we will detail the causes and impact of analytical variability in task-based fMRI, as well as various research questions that it opens up in neuroimaging research.

5.2 Analytical Variability in fMRI

Analytical variability refers to all variability resulting from the protocol applied on the already obtained raw fMRI data. Sources of variability in neuroimaging include the variability related to the variety of possible choices in fMRI pipelines, as well as the variability caused by technical conditions (software package, operating system) in the analysis.

In this section, we will detail the reasons for the existence of analytical variability in fMRI and the existing sources of variability, both related to pipeline operations and software/operating system conditions.

5.2.1 Why does Analytical Variability Exist?

In chapter 2, we have given a detailed description of processing steps which may be applied on fMRI data, why these steps may be applied - or must be applied - on the data, and the variety of choices which could be made by researchers. These include choices regarding the application or not of a certain operation on the data, a choice in parameter value or algorithm, or a choice in order of the operations.

In the development of fMRI pipelines, this large amount of methodological choices and the lack of gold standard led to a large range of possibilities in terms of pipelines considered acceptable for a same analysis, with limited ways to distinguish between them in term of quality.

Attempts to improve the quality of pipelines and methods used also lead to new processing possibilities, adding to the variability in terms of pipelines used over time in the literature.

One possibility in terms of reduction of variability is the optimization of pipelines. First, it limits the number of options for given processing operations, reducing the variety in terms of possible outcomes, and may help converging towards acceptable standards. Optimization of pipelines has notably been studied for various situations in [145, 24, 43].

However, multiple obstacles appear when considering the optimization of pipelines as a way to reduce analytical variability. First, it is not always clear whether a processing choice will be better than another for data processing. Processing options are used to correct problems existing in the raw data, but they may also be responsible for other issues. An example is the use of motion correction regressors 2.2.2. When adding motion regressors, sensitivity can be improved, on one hand, because of the reduction of within and between-subject variability [52, 98]. However, on the other hand, it can also be reduced in case of motion correlated to task [18]. A same problem may also be addressed differently by two different, independent processing operations, for example slice-timing may be partially corrected by the inclusion of temporal derivatives of the HRF [73] instead of using slice-timing correction.

Another problem is that optimal processing choices may depend on the data and what we want to do in the analysis. For example, slice-timing correction is less useful with small time of repetition (TR). Another example is the choice of different levels of smoothing, which may depend on what type of effects we may want to detect at statistical inference (lower or higher effect size, bigger or smaller size of the activated regions).

Variability in research choices may also arise from software conditions. Researchers often do not define by themselves the whole set of parameters in their analysis but use the settings given by the software packages, leading to software-driven constraints in research.

5.2.2 Variability in Pipeline Parameters

In this subsection, we will give an overview of some processing parameters for which there is a multiplicity of options, and how there impact have been studied. The possibilities in processing options detailed here are summarized in Table 5.1.

Preprocessing

Preprocessing consists in applying a set of operations in order to correct for artifacts, as well as applying spatial registration in order to have multiple subjects aligned to a same reference space. Researchers can choose to perform or not specific operations, such as slice-timing correction or despiking [20]. For operations such as registration or realignment, a number of options may be chosen [67] regarding various factors, including the type of transformation, optimization

Processing Step	Operation	Options
Preprocessing	Despiking	Presence or Absence
	Slice-Timing Correction	Presence or Absence
	Motion Correction	Realignment estimation techniques options Reslicing options Interpolation options
	Spatial Registration	Transformation estimation techniques options Choice of Template Reslicing options Interpolation options
	Smoothing	Smoothing kernel FWHM value
First-Level Analysis	Haemodynamic response	Modeling (gamma HRF, double gamma HRF, FIR model, CBS)
	High-pass Filtering	Yes/No
	Autocorrelation Correction	Yes/No
	GLM Modeling	Presence/Absence of temporal derivatves Number of Motion Regressors (0, 6, 12, 24)
Second-Level Analysis	GLM Modeling	Modeling of within- and between-subject variance (ordinary least squares, generalized least squares)
Statistical Inference	Type of Inference	Clusterwise/Voxelwise
	Thresholding	Corrected/Uncorrected
	Method	Parametric/Non-parametric

Figure 5.1 – Overview of some possible sources of variability in processing operations at each processing step.

algorithms, reslicing or interpolation, with various degrees of consensus regarding what are the best options for each. registration can be done using different templates, as it is done in [20].

Smoothing can be applied on the data with different possible levels of smoothing, defined by the FWHM of the smoothing kernel. At statistical inference, a lower level of smoothing will be associated to the detection of smaller regions with higher effect size, while a higher level of

smoothing will lead to the detection of larger regions. Smoothing level also has an effect on conservativeness when using Random Field Theory for statistical inference [72].

First-Level Analysis

As for preprocessing, first-level analysis contains optional steps which can be performed to correct for artifacts in the data, which researchers can choose to use or not, and for which they can choose specific paramter values. This include high-pass filtering, for which the high-pass filter value can be chosen by the researcher. Also, correction for temporal autocorrelation can be applied using different algorithms.

Another source of variability is the GLM modeling. As said in section 2.2, different models can be used for the haemodynamic response, including haemodynamic response functions using classical or double gamma functions, as well as Finite Impulse Response Models or Constrained Basis Sets. With the HRF, temporal derivatives of the HRF may be included in the model, to account for the issue of potential delays of the haemodynamic response within a subject. Also, motion regressors and their squares, derivatives and squares of derivatives may be included in the model in order to account for motion effects in the data.

Second-Level Analysis and Statistical Inference

One source of variability at group-level analysis is the type of modeling for between and within-subject variability. As said in 2.3, different methods can be used depending on considerations regarding within-subject variability. If we consider it to be equal, simple ordinary least squares can be used. Otherwise, generalized least squares have to be used, with a variety of possible algorithms.

Finally, there are multiple options for statistical inference. Voxelwise and clusterwise inference can be used, there are various possibilities for thresholding, and non-parametric methods may also be used. In the case of Random Field Theory, as we said previously, low levels of smoothing can result in conservativeness of the analysis [72].

Order of Operations

At the subject-level, the order of operations can also be done in multiple ways [22]. An example of this is registration, which may be performed before subject-level statistical analysis (as done with SPM) or after (as done with FSL). When done before, it is applied on the temporal sequence of 3D images of the BOLD signal. When it is done after first-level analysis, it is applied on the contrast maps.

Combination of Parameter Variations

The multiplicity of options leads to a very important number of potential pipelines, with potential differences in results across pipelines for a same research question. [20] used a number of different processing options on the same dataset to assess these differences in results depending on pipeline differences. The number of processing options resulted in 6,912 pipelines to obtain unthresholded results, and the application of 5 thresholding methods resulted in 34,560 thresholded maps. Figures 5.2 and 5.3 show the variability in results overall, as well as the variability associated to specific analysis parameters in different regions of the brain.

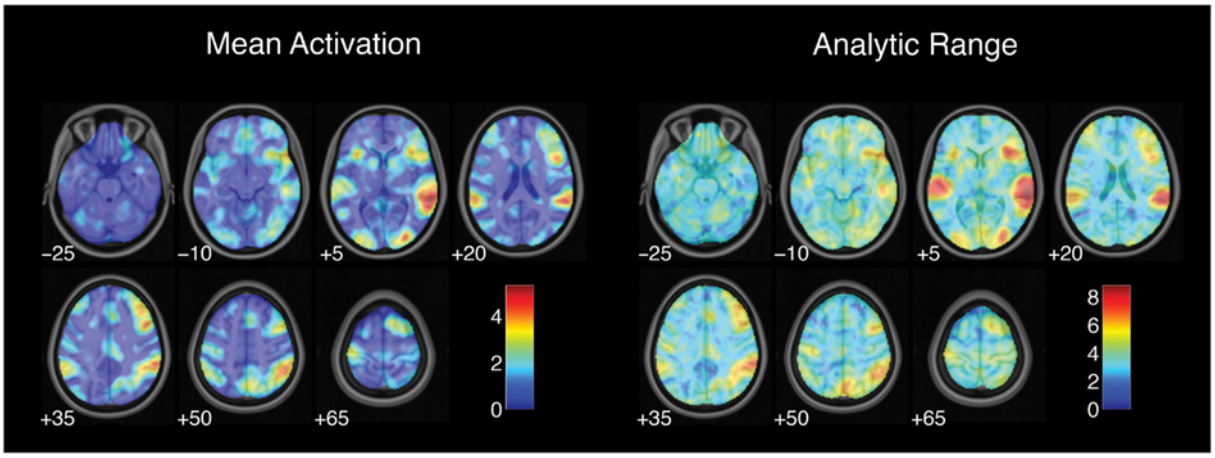


Figure 5.2 – Means (left) and ranges (right) of statistical Z-values obtained at each voxel across results obtained for all the different pipelines generated in [20]. 6,912 pipelines were applied on the same dataset from a study of response inhibition. Pipelines differed on a set of pre-defined parameters for preprocessing (despiking, slice-timing correction, spatial normalization, spatial smoothing) and model estimation (normalization-modeling order, high-pass filtering, temporal autocorrelation correction, run concatenation, model basis set and head motion regressions). Results show a high correlation between mean activation and analytic range (range of statistical Z-values) across voxels, with a larger variability in zones with high mean activation.

Another study which shows the variability in results which can arise from pipeline differences is [14]. 70 teams were asked to test nine ex-ante hypotheses on the same dataset, each using their own pipelines, with consistent results for some of them and strong variations across teams for others.

5.2.3 Variability in Software Conditions

Variability in Software Packages

We have shown how specific parameter variations can lead to variations in results. In practice, one of the main cause for these parameter variations is the use of different software packages,

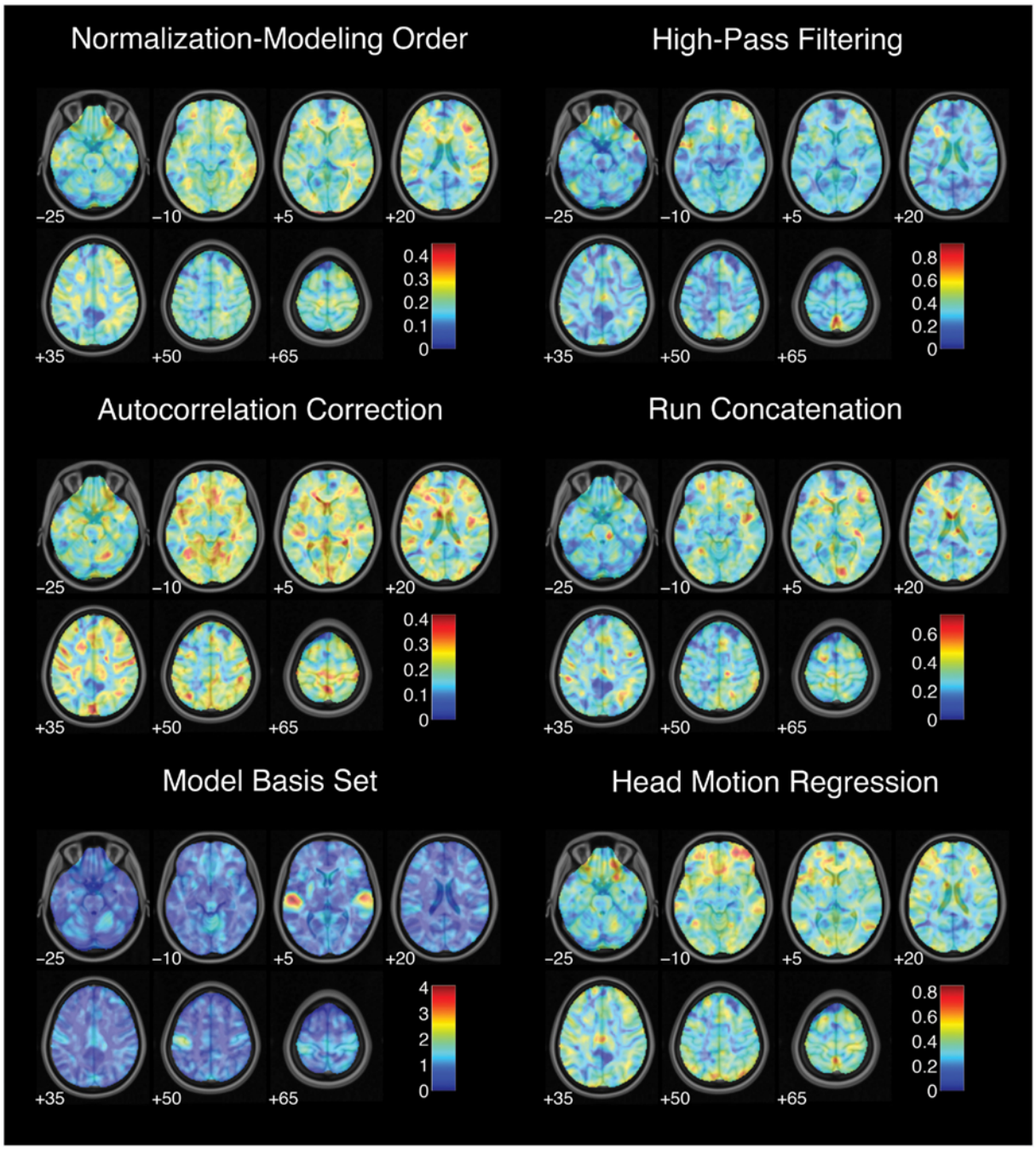


Figure 5.3 – Ranges of statistical Z-values obtained at each position of the brain over pipelines with a single parameter variations and other parameters fixed, for various first-level analysis parameters, in results from [20]. The results obtained show that the analytic range differ highly in term of position and intensity across parameters, notably with large variability associated to the model basis set in zones of high mean activation.

which is a common source of variability in general in neuroimaging, including outside of functional imaging [66]. Three software packages (SPM, FSL, AFNI) cover about 80% of fMRI research [20]. Each of these packages perform certain steps of processing and analysis of the data using different methodological choices or implementations.

In task-based fMRI, differences across software packages include presence or absence of certain processing options, default settings, differences in algorithms used to perform certain operations. Order of operations may also be done differently, for example FSL performs registration after first-level analysis by default. For these reasons, using a software instead of another to do an analysis leads to an important number of differences between processing pipelines used.

[15] performed a group analysis usingAFNI, FSL and SPM, finding substantial variability between software packages in T-statistic values and location of activations, with inter-software differences measured with Dice coefficient (Fig. 5.4), Bland-Altman plots and Euler Characteristics.

Another study, [117], used multiple variants of pipelines in the three software packages to perform analyses, finding consistency in activation patterns but also differences in AFNI likely due to the use of motion regressors (Fig. 5.5).

Variability in Operating Systems

Finally, one major problem regarding variability is the potential impact of the computing platform. While software conditions cause variability related to the use of different parameter values, operating system differences can lead to numerical differences in the analysis, potentially having meaningful effects on the results obtained [84, 85]. [58] studied the impact of operating system when performing an analysis with the same processing steps, for different neuroimaging problems (including resting-state fMRI analysis but also subcortical tissue classification and cortical thickness extraction), finding differences in fMRI analysis in particular in independant component analysis results across operating systems.

5.3 Research Questions related to Analytical Variability

We have seen in chapter 3 that lack of robustness is an issue, notably because of the analytical flexibility resulting from the variety of possible analysis choices in research, and in neuroimaging in particular. Here, we include a list of open research questions which are related to analytical variability.

5.3.1 Variability in the final results

The clearest impact of analytical variability is the variability in results which can occur when using different pipelines. [14] gives a practical example of researchers which find differences in

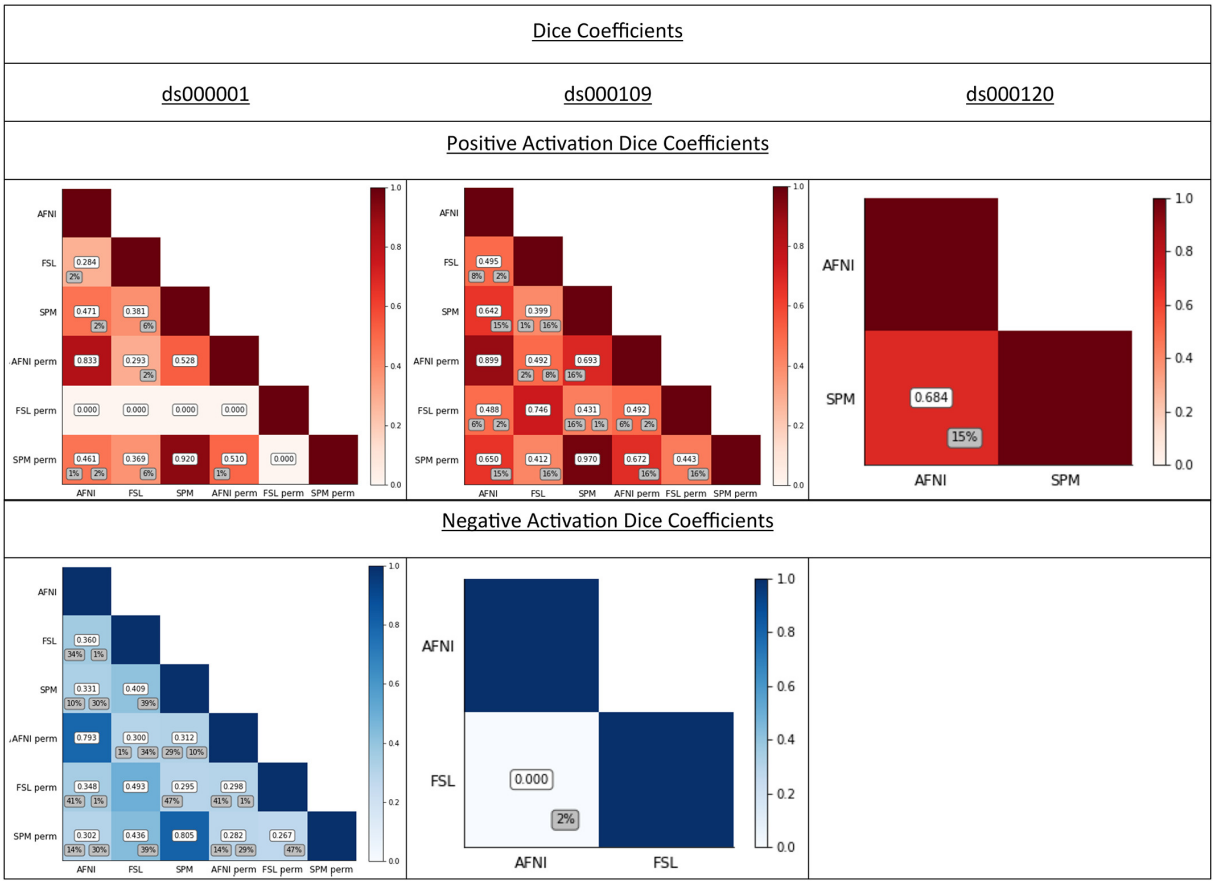


Figure 5.4 – Dice coefficients for pairs of positive activation maps and pairs of negative activation maps, obtained with different pipelines on the same dataset, for three different datasets, from [15]. Analyses over these datasets were performed using pipelines with SPM, FSL and AFNI, with and without nonparametric permutation test. Dice coefficient are mostly under 0.5 (with 0 for FSL with a permutation test on the first dataset because of the absence of activation), suggesting a large disparity in activation results across software packages. Pairs of analyses with the same software with and without permutation test show stronger overlap between activation maps, with Dice coefficient closer to 1.

results with different usual pipelines: 70 teams of researchers were asked to perform an analysis each with their own pipeline. In this situation, it is not known what the variability in results can be attributed to, because of the large number of differences across pipelines.

However, this study shows that, while there were strong variations in results in some cases, there also was relative consistency in other cases, suggesting that the converging results can be used to build a consensus in some situations. Using an important number of pipelines allows to see whether the variability in results across pipelines in a given situation is important or not, and whether a consensus can be reached. [20] observed the variability in results obtained with pipelines varying on a set of given parameters values, in various conditions for the combination

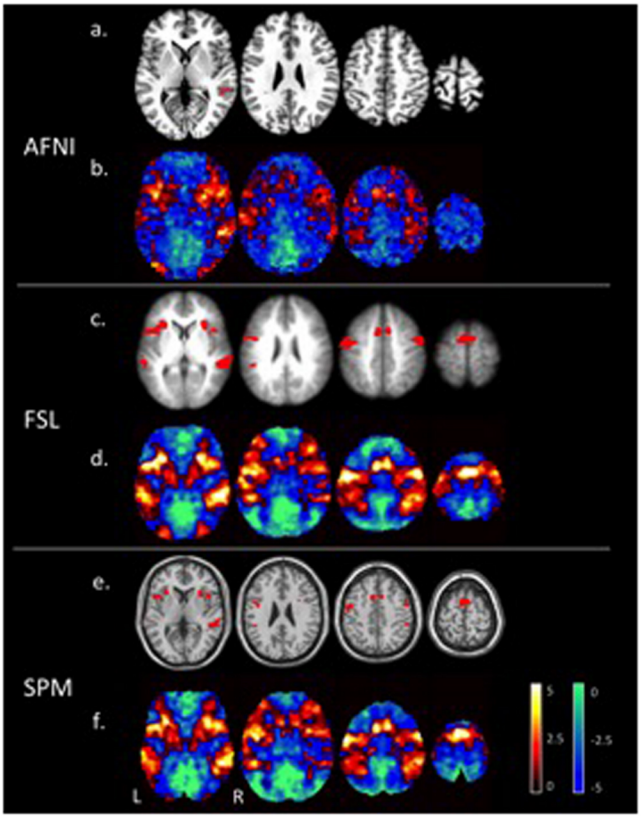


Figure 5.5 – Results for one of the variants of group-level analysis for a tone counting task contrast, from [117], for each of the three software packages used in the study (SPM, FSL and AFNI), thresholded with a $p < 0.05$ FWE corrected cluster-wise threshold and unthresholded. Results for the analysis with AFNI show lower activation than with SPM and FSL, although the patterns of activation are similar. This may be caused by the default presence of motion regressors in AFNI, which can lead to lower sensitivity and is not included here in analyses with SPM and FSL.

of other parameters, by observing variations in activation strength across results obtained with different pipelines.

5.3.2 Analytical Flexibility

A direct consequence of this variability in results is the risk of analytical flexibility. As said in chapter 3, a large number of pipelines can lead to the possibility of having at least one of them mistakenly giving a false positive result, even if this pipeline is not invalid in itself, as shown in [20]. Same as for result variability, the repetition of research results with other pipelines can be used to have more confidence in these research findings.

5.3.3 Validity of the Final Results

Because of the large number of pipeline parameters for which methodological choices for each operation are deemed to be acceptable, there can be processing choices which are not usually considered unacceptable and are commonly used, but actually increase the chance of finding false positive results in a study in certain situations. [40] shows the inflated risk of false positive rates using clusterwise inference under certain circumstances, suggesting that false positives in this case were obtained because of an existing effect of a specific combination of parameter choices. When invalidity linked to pipeline parameters is observed, these parameters should be avoided in further studies. Optimization of pipelines and procedures of validation is a possible way of trying to restrict choices in order to avoid those leading to invalidity in results. However, there are limits due to the difficulties in determining what processing choices are better than the others, as explained in 5.2.1.

5.4 My Work

We saw that existing work regarding analytical variability in fMRI were limited to the case where researchers use a complete pipeline with identical subject-level processing on all subject data. They showed that pipeline differences across analyses could cause important changes in results obtained across analyses because of analytical variability [20, 14, 40].

In the case where researchers want to combine subject data processed with different subject-level pipelines within a same group study, they may want to know whether or not these pipeline differences have an effect on the results obtained. For this reason, we perform between-group analyses using subject data processed differently, and observe the effect of the variability related to differences in processing on the results in this situation.

Similar situations have been studied in the case of other sources of variability in data, in various fields, including neuroimaging. In particular, for technical variability, methods have been developed to estimate and remove the inter-scanner or inter-site effects when combining data from different sites [47]. Some of these methods have been adapted from existing methods in other fields, for example methods to remove batch effects and other sources of unwanted variations in genomics [55]. However, such approaches do not exist yet to our knowledge for analytical variability.

Frameworks for the optimization of pipelines in neuroimaging studies have considered situations where, for group studies, the optimal subject-level processing may vary depending on the subject, resulting in group analyses using different subject-level pipelines [145, 139]. However, they also address a situation different to ours: besides the different objectives, these study consider a situation in which a researcher may choose different pipelines and is looking for an optimal choice, whereas researcher who combine pipelines processed differently do not have choices

over the processing options already applied on the data they want to combine.

For neuroimaging group studies, the case which have been addressed regarding analytical variability is the situation where an analysis is done using a complete pipeline (which performs a same subject-level processing on all subject data and second-level analysis) [20, 14, 40]. These studies repeat the analysis multiple times, each time with a different complete pipeline, to compare the results across pipelines, but there are no subject-level processing differences within a same iteration of the analysis. This is not the case in our situation, where subject-level processing differ across subjects within a same analysis. The difference between both cases is described on Figure 5.6.

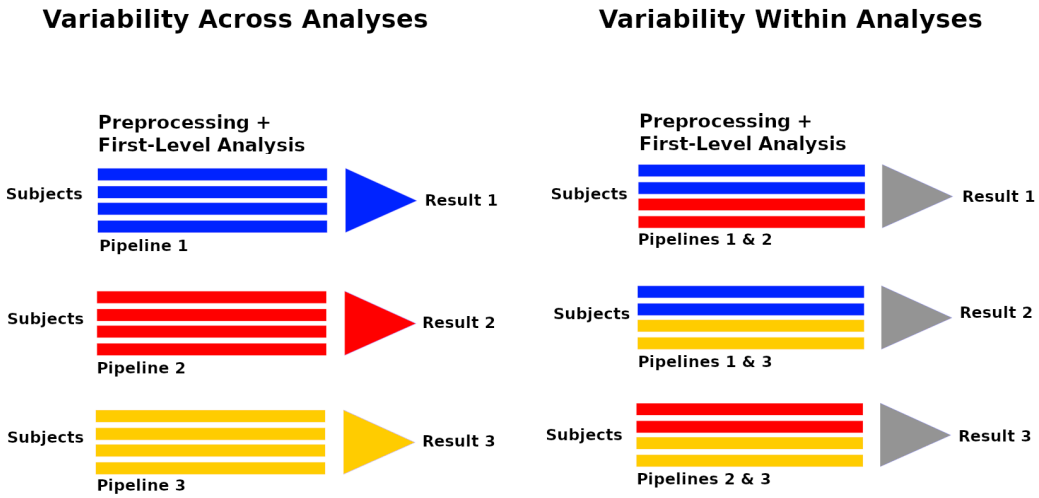


Figure 5.6 – Description of the differences between situations explored in previous works related to analytical variability, such as [20, 14] (left) and in our work (right). In existing works, a same analysis is repeated multiple times, with a same processing applied on all subject data within each analysis, and processing differences across analyses. Results obtained for each analysis can then be compared. In our work, we perform a between-group analysis combining subject data processed by two pipelines, and repeat this for multiple pairs of pipelines. The between-group analysis is identical for all pairs of pipelines, and results obtained for each pair can be compared.

Applying a same subject-level processing on all subjects within a study is the standard way of doing fMRI studies, and it is common to have multiple studies performing experiments to address a same research question with processing differences across them. On the other hand, combination of subject data within a same study – with the potential existence of processing differences between them – has been made possible by recent developments of data sharing

practices in neuroimaging. This may explain why the first situation has been addressed in existing works and not ours. We hope that – thanks to research understanding how analytical variability can be dealt with – in the future combination of datasets processed differently will be made possible and will be commonplace to enable larger sample sizes.

Chapter 6

First Contribution: Estimation of the Validity of Group Analyses Using Subject Data Processed Differently

Part of the work done for the contribution presented in this chapter has been published for ISBI 2022 [131]. Another article presenting the rest of the contribution done is currently in work.

6.1 Introduction

Task-based functional Magnetic Resonance Imaging (fMRI) studies the activation of brain regions during the realization of a task. The main form of fMRI is Blood-Oxygen Level Dependent (BOLD) fMRI, which uses MRI to measure, at every position of the brain, a BOLD signal whose variations in time are related to brain activity. For a given subject, the data resulting from an MRI session is a temporal sequence of 3D brain images. In fMRI studies, these data are acquired, and then preprocessed and analyzed at the subject and group-level.

Although fMRI has been a useful tool to provide information about the roles of the different regions of the brain, multiple concerns have been raised over the last few years regarding research practices and factors which might impact the reproducibility of fMRI studies [20, 15, 146]. One of those factors is the overall low statistical power of fMRI studies: having low sample sizes makes it harder to find true positive results, hence increasing the likelihood that any positive result is false [19]. While the sample sizes in fMRI studies have been increasing over the past few years (in 2015, the median was at almost 30 subjects for single-group studies [123]), there is still a crucial effort that has to be done in order to increase statistical power.

There are multiple ways to overcome this lack of reproducibility induced by low sample sizes. For example, meta-analyses can be used to combine results from multiple studies (group-level

statistical maps in our case) to observe converging results [134]. However, there are several limitations to this method, notably due to the existence of the publications bias[78]. Another possibility is to take advantage of data sharing: today, with the importance given to open science [125] and the development of research infrastructures [124, 64], more and more neuroimaging data from various studies are made available to the scientific community. These datasets includes subject data from multiple different studies, which can be used and combined in new studies. Being able to combine data would allow us to have larger sample sizes per study and thus more robust results.

While most studies based on data reuse currently focus on raw data, we expect that in the future more and more processed datasets will be made available. Sharing of processed data may appear because of privacy reasons, and also in order to avoid having to perform the subject-level processing each time someone wants to reuse the subject data. FMRI studies require multiple steps of processing on the data, both at the subject-level (preprocessing of the raw fMRI data to make it fit for statistical analysis, and first-level analysis for each subject) and at the group-level (second-level statistical analysis using the subject data resulting from first-level analysis).

Within a study, there are many factors of variability which may have an effect on the analysis. Some of these are unwanted factors of variability on which the results of a study may depend. Sources of variability include acquisition instruments, acquisition protocol, and differences in the processing and analysis protocol [48]. Here, we focus on the variability resulting from the processing and analysis protocol used on the data, which is called analytical variability. This type of variability has been studied in various cases (for example the variability related to software and software versions in [15]). At each processing step within fMRI studies, multiple methodological choices are available: choosing to do certain operations or not, choosing their order, choosing certain parameter values for a specific operation, choosing a software package or another which may use a different tool to perform an operation [20]. A given series of operations that will be performed on the data is called a 'pipeline'. Pipelines can be used to perform all the steps of the analysis, or only parts of it: for example, subject-level pipelines only perform preprocessing and first-level analysis on the subject data.

The variety of possible methodological choices leads to a large amount of possible pipelines for a same analysis. It has been shown in multiple studies [56, 66] that variability in analytical choices in neuroimaging studies can lead to differences in results when doing a similar experiment with different methods of analysis, in particular in fMRI when using different pipeline [15, 20, 58]. In [14], 70 teams performed similar analyses, each with a different pipeline, and found substantial differences in the results obtained across teams. Other studies have explored the variability resulting from specific aspects of the processing and analysis protocol, including operating system [58], differences of software and software versions [117, 15](which have different implementations to perform various operations), as well as the choices regarding the processing

steps applied on the data [20].

The question of whether analytical variability makes it impossible to combine subject data for between-group analyses in these conditions is still widely unanswered. Previous studies have focused on how analytical variability affects the reproducibility of existing results in neuroimaging, by using different pipelines to complete a similar analysis in which the processing applied on all subject data is the same, and comparing the results obtained across pipelines using different processing pipelines. Specific frameworks for the optimization of pipelines by estimation of performance metrics associated to reproducibility [139, 89, 145] have been developed to address the issue of analytical variability in this situation. Notably, solutions using different subject-level processing pipelines depending on the subject have been suggested in this context [25]. Here, we performed between-group analyses using subject data processed with different pipelines to study the compatibility between data in these conditions.

As mentioned, we may expect to see a growing proportion of processed data among shared subject data in the future. If we want to use these data, coming from different sources, for between-group studies, it would be necessary to ensure that the processing differences will not impact the results of between-group analyses and give us something different from what we would have had with identical processing for each subject. In this study, we focus on how the analytical variability in subject-level pipelines can impact the results of between-group studies.

In order to assess the validity of between-group studies combining data that were processed differently at the subject-level, we carried out a series of between-groups analyses under the null hypothesis, using data from the Human Connectome Project [152], preprocessed differently. We processed the raw data from all subjects with a set of preprocessing and first-level analyses that are often used in the literature, with subject-level pipelines that are typically used in the literature. These subject-level pipelines, which consist of the same steps and are used under the same practical conditions, only differ on a set of predefined parameters. In the following sections, the term 'pipeline' will be used to designate subject-level pipelines only.

False positive rates were used to assess the validity of between-group analyses with different processing pipelines. The interest of having bigger sample sizes is to increase the chance of detecting true positive results: having inflated false positive rates when using different pipelines, compared to the expected false positive rate under null hypothesis, would suggest that it is also easier to find false positives when combining data processed differently, making the larger sample sizes useless.

Having a specific pipeline for each group data is not the only situation where we may have multiple subject-level pipelines within an analysis: we may also have group studies where there are multiple pipelines used within each group. The situation where pipelines are confounded with groups can happen, for example, when using subject data coming from a specific dataset for each group, and each dataset is associated to a specific processing pipeline.

6.2 Material and Methods

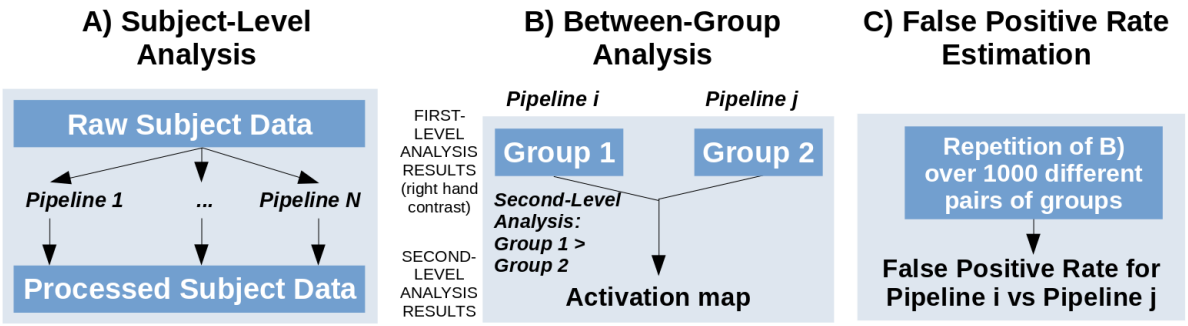


Figure 6.1 – Overview of the method: subject-level analysis (A), between-group analysis (B) and estimation of the false positive rate (C).

This study was performed using data from the Human Connectome Project. Written informed consent was obtained from participants and the original study was approved by the Washington University Institutional Review Board. We agreed to the Open Access Data Use Terms available at: `https://www.humanconnectome.org/study/hcp-young-adult/document/wu-minn-hcp-consortium-open-access-data-use-terms`.

The goal of the study was to test the validity of between-group analyses using subject data processed with different pipelines. To this aim, we performed between-group analyses under the null hypothesis (no between-group difference). All significant differences detected were therefore false positives, and we used the detection rate as an estimate of the false positive rate (which under the null hypothesis is expected to be equal to the nominal false positive rate). In this study, we estimated empirical false positive rates obtained when comparing two groups of subjects, each processed with a different pipeline, and observed whether or not they diverged from the nominal false positive rate. Subjects in each group were randomly drawn from the Human Connectome Project (HCP) 1200-subject dataset [152, 106, 44, 137, 163].

The steps performed in order to estimate this false positive rate, which will be detailed in the following subsections, are presented on Figure 6.1: first, the raw fMRI data was processed through each subject-level pipeline, for all subjects (section 2.2). Then, for each pair of pipelines that we wanted to compare, we performed a between-group analysis on the first-level analysis results, with subject data processed with one pipeline for one group and with the other pipeline for the other group (section 2.3). This group comparison was repeated 1,000 times in order to estimate the empirical false positive rate (section 2.4).

All the scripts used to perform the study (subject-level processing and analysis, group-level analysis, false positive rate estimation) are available at `https://github.com/Inria-Empenn/pipelines`.

6.2.1 Material

We used unprocessed fMRI data associated with the motor task and structural data for all subjects from the HCP 1200-subject dataset [152] who had completed this task ($n = 1080$). Multiple preprocessing and first-level analyses were performed on the fMRI data (see section 2.2).

6.2.2 Subject-Level Pipelines

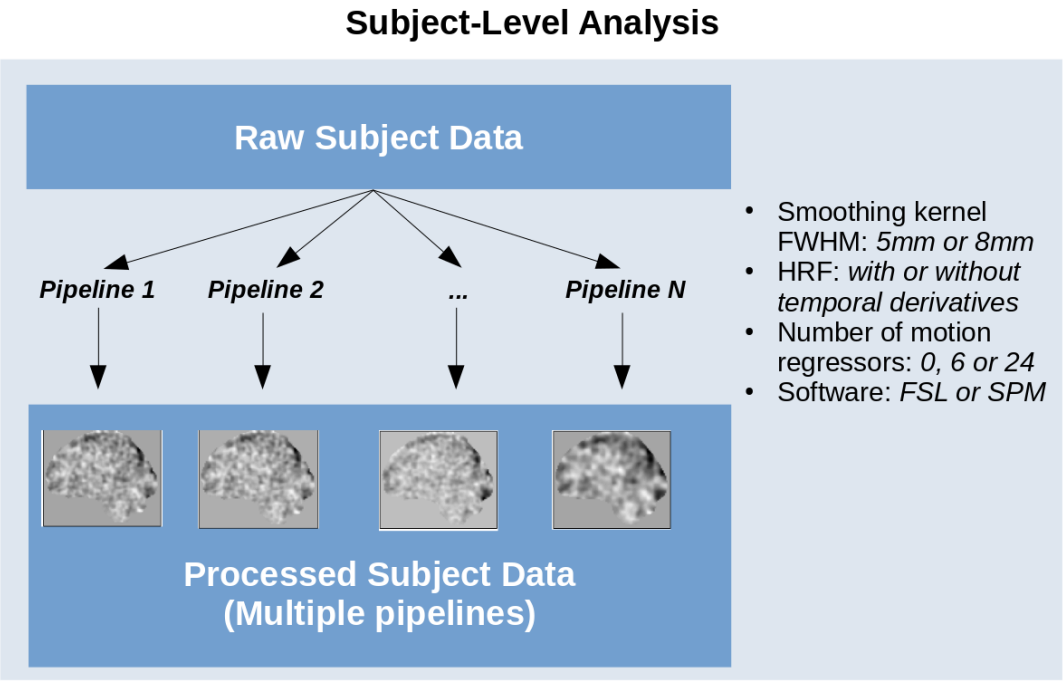


Figure 6.2 – Description of the subject-level analysis. Each subject data is preprocessed and analyzed with all different subject-level pipelines. The parameters that vary across the pipelines are the smoothing kernel FWHM value, the number of motion regressors included in the GLM model, the presence or absence of temporal derivatives of the HRF and the software package.

All subject data were processed through multiple pipelines, which carried out preprocessing and first-level analysis, as shown on Figure 6.2. Preprocessing consisted of the following steps for all pipelines: spatial realignment of the functional data, coregistration of realigned data towards the structural data, segmentation of the structural data, non-linear registration of the structural and realigned functional data towards a common space and smoothing of the registered functional data. For each pipeline, we selected the contrast corresponding to the right hand in the motor task.

The subject-level pipelines used for the analysis were performed using SPM12 r7771

(RRID: SCR_007037, [118]) with Octave 5.1.0 (RRID: SCR_014398, [38]), and FSL 5.0.1 (RRID:SCR_002823, [80]). All analyses were performed under Debian 10.6. The operations performed on the data for subject-level analysis were done in the default order for each software package: in SPM, they were carried out in the order described previously. For SPM, apart from the varying parameters defined below, each pipeline used the default settings; for FSL, we defined pipelines that were as close as possible to the pipelines defined for SPM. high-pass temporal filtering value was set to 128s to have the same value as SPM. The order of operations performed on the subject data was similar, except for spatial registration which was done on the contrast maps after first-level analysis as traditionally done with FSL.

We looked at the following set of parameters:

— Smoothing kernel: Full-Width at Half-Maximum (FWHM) was equal to either 5mm or 8mm.

— Number of motion regressors in the General Linear Model (GLM) for the first-level analysis: 0, 6 (3 rotations, 3 translations) or 24 (6 head motion regressors + 6 derivatives and the 12 corresponding squares of regressors).

— Presence or absence of the temporal derivatives of the Haemodynamic Response Function (HRF) in the GLM for the first-level analysis.

In total, those combinations provided a set of 12 different subject-level pipelines for each software package (2 FWHM $\times$ 3 numbers of motion regressors $\times$ 2 HRF).

6.2.3 Between-group Analyses

We performed a between-group analysis comparing two groups of 50 subjects, where two pipelines were selected and respectively applied to all subjects of each groups, for multiple pairs of pipelines. The 50 subjects in each group were randomly sampled without replacement, uniformly among the 1080 subjects.

Figure 6.3 provides an overview of the between-group analysis: using the contrasts resulting from first-level analysis for these subjects and pipelines, we looked at the second-level contrast corresponding to the between-group difference in means. We performed a one-tailed t-test with unequal variance using a voxelwise $p < 0.05$ FWE-corrected threshold to detect whether there was any significant between-group difference.

Since there were 12 different pipelines for each software package, we performed a total of 144 between-group analyses per neuroimaging software package (SPM, FSL). In addition to the between-group analyses investigating the compatibility between different pipelines (that we will refer to as "different pipeline analyses" in the following), we also checked the validity of the tests when using the same pipeline for both groups (referred to as "identical pipeline analysis" in the following).

In order to have consistent second-level software and analysis conditions for all between-group

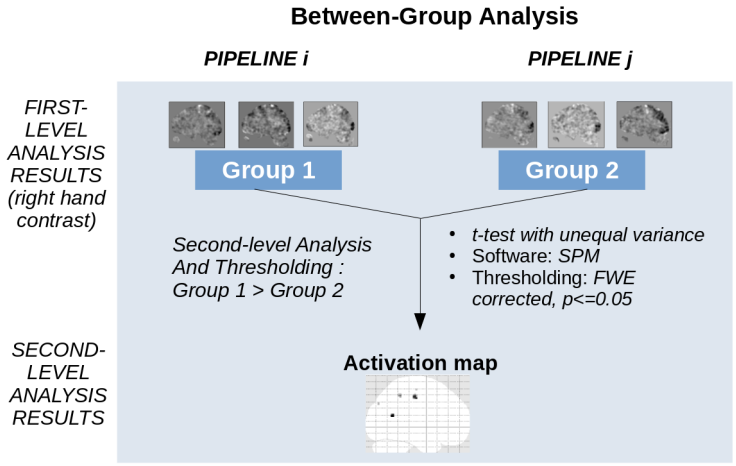


Figure 6.3 – Between-group analysis using two different subject-level pipelines: a between-group analysis for the right hand contrast resulting from the first-level analysis is performed between Group 1 (with subject data preprocessed and analyzed with pipeline i) and Group 2 (with subject data preprocessed and analyzed with pipeline j), using a one-tailed t-test with unequal variance and a $p < 0.05$ FWE-corrected thresholding to obtain an activation map. Both groups are composed of 50 subjects, with the 100 subjects being all different and sampled randomly, uniformly among the 1080 subjects.

analyses, all second-level analyses were performed with SPM.

6.2.4 False Positive Rates Estimation

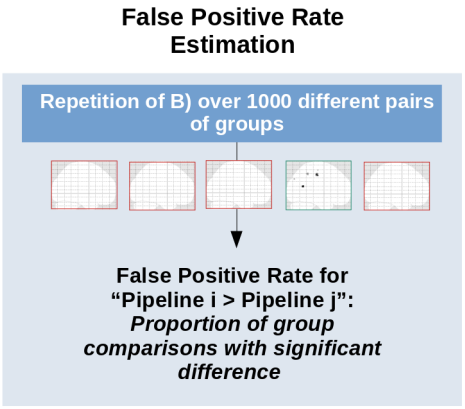


Figure 6.4 – Estimation of the false positive rate for a given pair of pipelines. The between-group analysis shown in Figure 6.3 is repeated 1000 times, giving between-group analyses with either presence (green) or absence (red) of significant between-group difference. The proportion of between-group analyses with at least one significant difference is the estimate of the false positive rate.

Under null hypothesis, there is a theoretical 5% chance to detect any significant difference in activation for a between-group analysis. Thus, for multiple between-group analyses, the proportion of analyses with at least one detected significant difference should converge towards this 5% false positive rate.

In order to have a sufficient number of analyses to observe the convergence of false positive rates, each between-group analysis with groups processed with different pipelines was repeated 1000 times with different groups, to estimate the false positive rates for each pair of pipelines. The empirical detection rate was the proportion of between-group analyses, over the 1000 repetitions, with at least one significant difference detected between two groups, as shown on Figure 6.4.

The set of 1,000 pairs of 50-subject groups used for the estimation of the false positive rate was the same for all pairs of pipelines.

6.3 Results

6.3.1 Analyses using the same pipeline

Figure 6.5 shows the false positive rates obtained for all cases of identical pipeline analysis, both for SPM and FSL. There is an identical pipeline analysis for each of our 12 possible pipelines, both in SPM and FSL. For all identical pipeline analyses, the false positive rates obtained were below the expected value of 0.05 for both software packages, ranging between 0.013 and 0.024 for SPM and between 0.012 and 0.019 for FSL (see Fig. 6.5). The exhaustive list of false positive rates for all analyses performed is displayed in the matrices on Figures 7.5 and 7.6 in the Appendix.

Results obtained for identical pipeline analyses can be used as a reference to be compared with the results obtained when using different pipelines.

6.3.2 Between-group analyses with different HRF models

The following subsections explore results obtained with different pipelines analyses, for which various types of observations were made in each case: false positive rates, statistical distributions and associated P-P plots. Default parameter values were defined for each parameter: absence of temporal derivatives of the HRF, 5mm for smoothing kernel FWHM value and 24 motion regressors. For false positive rates, all different pipeline analyses where both pipelines differ on a parameter, with value 1 on the first group pipeline and value 2 for the second group pipeline, are subsequently referred to as "parameter value 1 > parameter value 2" (for example, "5mm > 8mm" for a difference in smoothing). For P-P plots and statistical distributions, the results shown on the related figures are obtained for analyses with default parameter values when the parameter is not mentioned: in this case, for example, "5mm > 8mm" refers to "no

SPM

	Smoothing, 5 mm		Smoothing, 8 mm	
	No derivatives	Presence of derivatives	No derivatives	Presence of derivatives
No motion regressors	**0.087**	**0.071**	**0.194**	**0.200**
6 motion regressors	0.019	0.024	**0.097**	**0.107**
24 motion regressors	0.020	0.015	**0.051**	**0.101**

FSL

	Smoothing, 5 mm		Smoothing, 8 mm	
	No derivatives	Presence of derivatives	No derivatives	Presence of derivatives
No motion regressors	0.014	0.012	0.013	0.015
6 motion regressors	0.013	0.013	0.015	0.016
24 motion regressors	0.014	0.016	0.019	0.017

Figure 6.5 – False positive rates obtained for identical pipeline analyses with SPM and FSL, for all possible sets of parameter values for number of motion regressors, smoothing kernel FWHM and presence or absence of HRF temporal derivatives.

HRF temporal derivatives, 5mm, 24 motion regressors > no HRF temporal derivatives, 8mm, 24 motion regressors".

Adding a temporal derivative to model the HRF was the least impacting of all three varying factors in both software packages. Figure 6.6 shows the false positive rates obtained for different pipeline analyses with pipelines differing on one parameter or combination of parameters, both under FSL and SPM, including analyses with different models of the HRF. The false positive rates are given for analyses using pipelines with different models of the HRF, same smoothing and same number of motion regressors, for all possible values of smoothing kernel FWHM (5mm, 8mm) and number of motion regressors (0, 6, 24). In each case (six in total), we observed the false positive rates obtained both left tail and right tail, as well as the one obtained for identical pipeline analysis (with the default value of no derivatives of the HRF in both pipelines) for

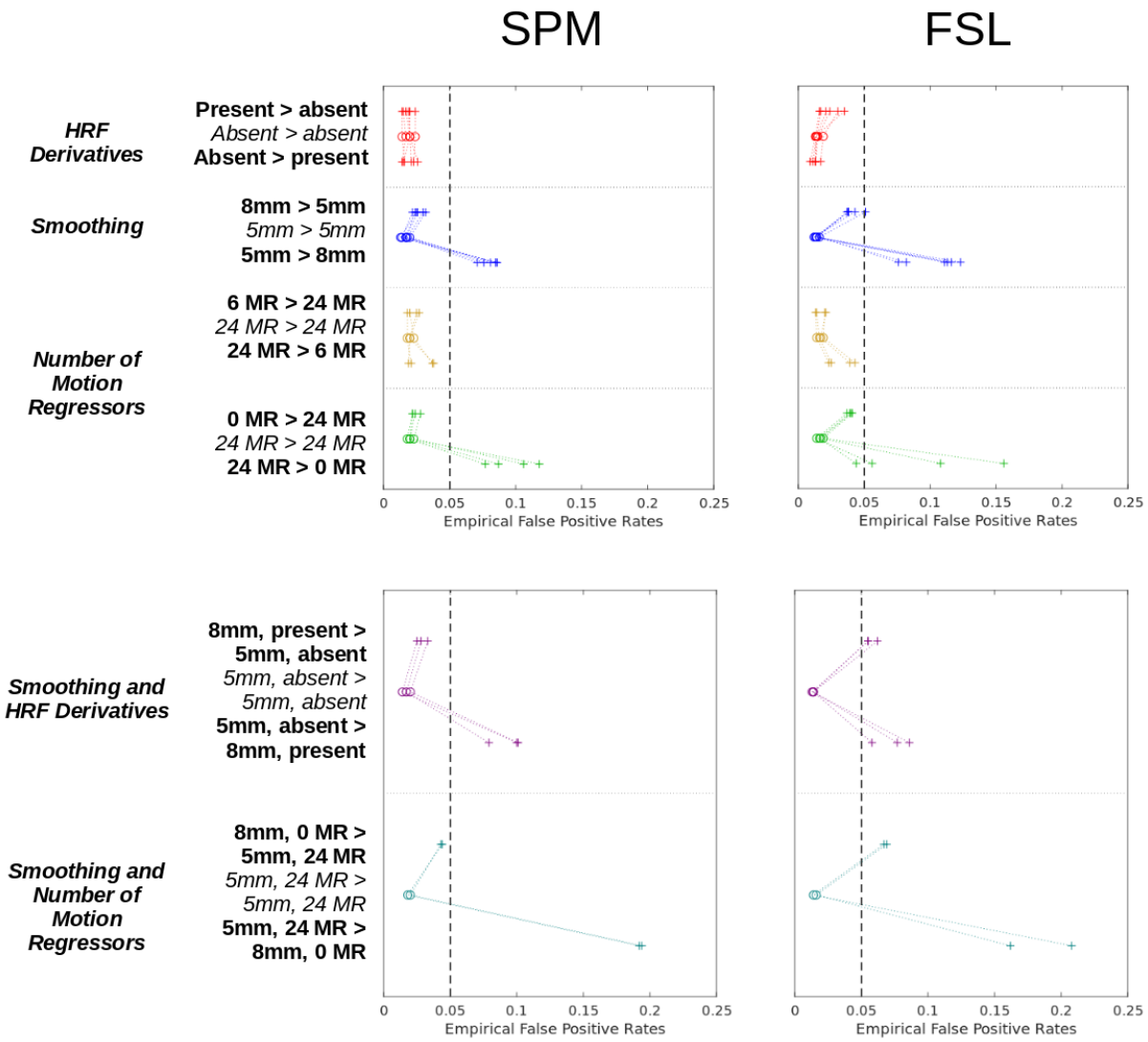


Figure 6.6 – Empirical false positive rates obtained for various between-group analyses with different or identical pipelines, in SPM (first column) and FSL (second column).
For each software package, we have :
-First row: for each specific parameter, given a default value and a variation, false positive rates obtained for all different pipeline analysis where the pipelines differed only on this parameter, in both directions (cross dots), compared to identical pipeline analyses where this paramater has the default value in both pipelines (round dots).
-Second row: similar results with two varying parameters instead of one.

comparison. All results obtained were conservative, with false positive rates reaching 0.024 for SPM and 0.035 for FSL.

In order to better understand how the variations in parameters affect the false positive rates,

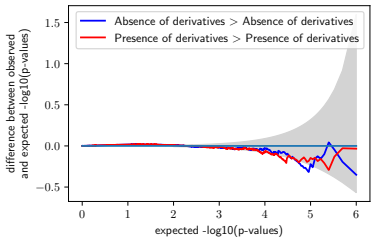
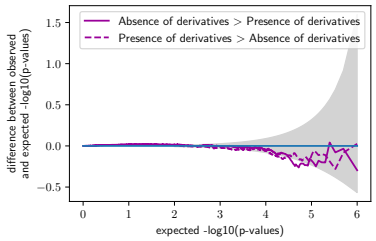
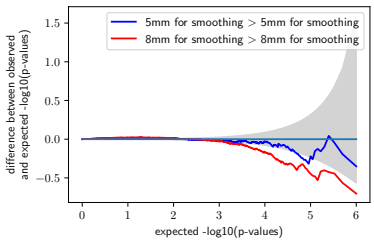
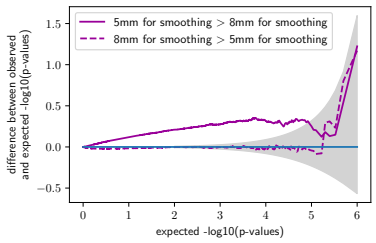
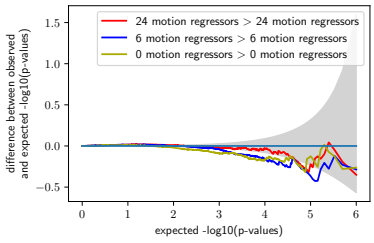
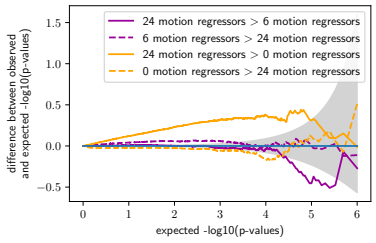
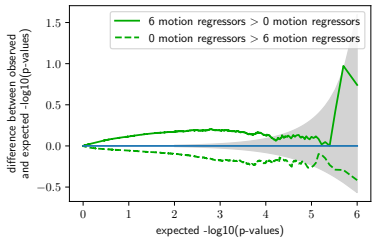
	Identical Pipelines	Different Pipelines
Effect of HRF		
Effect of Smoothing		
Effect of Motion Regressors		 

Figure 6.7 – variants of P-P plots for distributions obtained with various analyses under SPM, against the expected distribution, with 0.95 confidence interval, with a single varying parameter for different pipeline analyses. The variations from usual P-P plots are the use of -log(p-values) instead of p-values (to have a more precise observation of the behavior in the right tail) and replacing the obtained -log(p-values) on the y-axis by the difference in obtained and expected -log(p-values).
When not indicated, the default parameters for the pipelines were: 5mm for the smoothing kernel FWHM, 24 motion regressors and no temporal derivatives of the HRF in the GLM.
Positive differences indicate invalidity whereas negative differences indicate conservativeness.

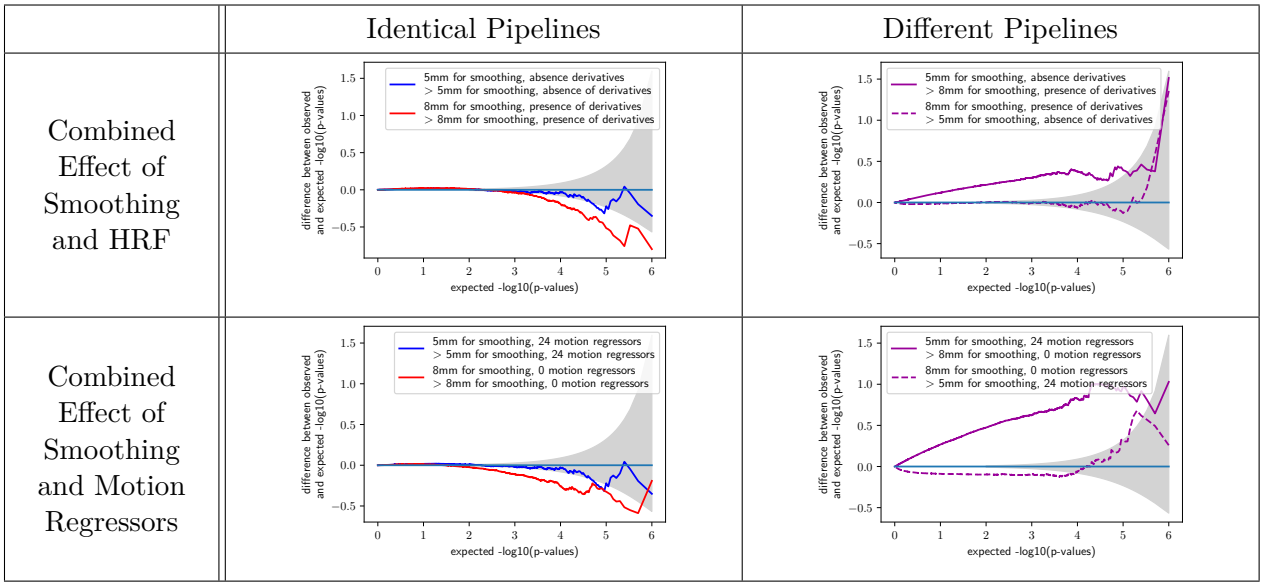


Figure 6.8 – variants of P-P plots for distributions obtained with various analyses under SPM, against the expected distribution, with 0.95 confidence interval, with a combination of varying parameters for different pipelines analyses. The variations from usual P-P plots are the use of -log(p-values) instead of p-values (to have a more precise observation of the behavior in the right tail) and replacing the obtained -log(p-values) on the y-axis by the difference in obtained and expected -log(p-values).
When not indicated, the default parameters for the pipelines were: 5mm for the smoothing kernel FWHM, 24 motion regressors and no temporal derivatives of the HRF in the GLM.
Positive differences indicate invalidity whereas negative differences indicate conservativeness.

for multiple between-group analyses, we observed the distribution of the statistical values over our 1,000 repetitions for voxels contained within the mask of the second-level analysis (For practical reasons, for each analysis, observations were made on a sample of 1,000,000 statistical values chosen at random uniformly from all statistical values across the 1,000 repetitions). Figures 6.9 and 6.12 show us, under SPM and FSL respectively, how the distributions obtained differed from the expected Student distribution for some of these analyses, including “absence of derivatives > presence of derivatives". For both software packages, there was no obvious deviation of the mean and variance compared to the expected theoretical distribution for distributions obtained for analyses using pipelines with and without temporal derivatives of the HRF.

For the statistical values in the distributions obtained with these analyses, associated p-values were also obtained and sorted in order to be compared to the corresponding expected p-values in the tails. The comparison was done using variants of P-P plots. Figures 6.7 show the P-P plots obtained for some of these analyses in SPM (including “absence of derivatives > presence of derivatives"), and Figures 6.10 the P-P plots for the same analyses in FSL. For

both SPM and FSL, in both directions, most values were within the 95% confidence interval (except for “absence of derivatives > presence of derivatives" with FSL where there was slight conservativeness).

6.3.3 Between-group analyses with different levels of smoothing

Smoothing was the most impacting of all three varying factors. Figure 6.6 shows the false positive rates obtained for the six cases of analyses with different levels of smoothing (in both pipelines, presence or absence of temporal derivatives of the HRF and 0, 6 or 24 motion regressors) with both softwares. For between-group analyses using pipelines with different smoothing (5mm and 8mm), in the direction “5mm > 8mm", the false positive rates were inflated both for SPM (ranging from 0.071 to 0.086) and FSL (ranging from 0.076 to 0.123). In the other direction, all false positive rates obtained stayed under the 0.05 theoretical false positive rate for SPM, and under or close to it in FSL. Compared to identical pipelines analyses, the false positive rates in both directions were always increasing. Observations in the tail on P-P plots (Fig. 6.7 and 6.10) show that between-group analyses using pipelines with different smoothing gave invalid results for both software packages in the direction “5mm > 8mm" (thus explaining the inflated false positive rates obtained for this combination of parameters), and values within the 95% confidence interval in the other direction (8mm > 5mm).

The behaviors observed on the P-P plots can be explained by the positive shift in mean and higher variance observed on the statistical distribution for “5mm > 8mm" (Fig. 6.9 and 6.12). These effects on mean and variance both caused an increase of statistical values in the tail, which explains the invalidity in this direction. In the other direction, however, values stayed within the 95% confidence interval because the shift in mean was negative, and the resulting decrease of statistical values in the tail compensated the increase caused by the higher variance.

6.3.4 Between-group analyses with different number of motion regressors

For analyses with different numbers of motion regressors, the amplitude of the effect depended on the parameter value chosen in both pipelines. Figure 6.6 shows the false positive rates obtained with both software packages for the four cases (presence or absence of derivatives, 5mm or 8mm for smoothing in both pipelines) with different numbers of motion regressors, for analyses with 24 motion regressors in one pipeline and 6 or 0 in the other. The biggest effect on the false positive rate was observed for between-group analyses with 24 motion regressors in one pipeline and none in the other. For both software packages, the between-group analyses for “0 motion regressors > 24 motion regressors” gave higher false positive rates than identical pipeline analyses – though always lower than the 0.05 theoretical false positive rate (Fig. 6.6). In the other direction (“24 motion regressors > 0 motion regressors"), we almost always obtained inflated false positive rates, ranging between 0.077 and 0.118 for SPM and between 0.044 and 0.156 for FSL. With

FSL, among all four between-group analyses corresponding to “24 motion regressors > 0 motion regressors", analyses with 5mm for smoothing in each pipeline gave the lowest false positive rates (0.056 and 0.044) and analyses with 8mm for smoothing in each pipeline the highest (0.156 and 0.108).

Analyses with 24 and 6 motion regressors gave valid results: the false positive rates were always under the 0.05 theoretical false positive rates. A slight increase in false positive rates can be observed for the between-group analyses “24 motion regressors > 6 motion regressors" compared to identical pipeline analyses.

For SPM, we observed shifts in mean and increase of variance on the obtained statistical distributions, similarly to what had been observed for smoothing (see Section 3.3) (Fig. 6.9). For the analyses using pipelines with 24 motion regressors and no motion regressors, the combined effects of variance and mean gave invalid results in one direction and valid results in the other, as we see on the P-P plots (Fig. 6.7), similarly to the results obtained when performing analyses using pipelines with different smoothing kernel FWHM. Results for 6 versus no motion regressors were similar but lower in both directions (slight invalidity in one direction, conservativeness in the other), explaining the lower false positive rates in both directions and suggesting a less pronounced effect of the increased variance in that case.

For FSL, we observe an increase in variance for the analysis “24 motion regressors > 0 motion regressors”, similarly to SPM, but no shift in mean (Fig. 6.12). While SPM gave strong invalidity and inflated false positive rate for the same analysis in the same direction and valid results in the other direction, with FSL, false positive rates and curves on P-P plots were very similar in both directions (Fig. 6.10). Also, the results observed on the distributions and P-P plots for analyses using pipelines with 6 and no motion regressors seem to indicate a shift in mean for FSL that was both less important and oriented differently compared to SPM. However, the distributions and P-P plots in these cases were observed on the between-group analysis with 5mm for smoothing and no derivatives of the HRF, which gave lower false positive rates than the analyses with 8mm instead. Therefore, these differences may not be representative of the real difference existing between results in FSL and SPM.

6.3.5 Combined effects of parameters

We observed the combined effects of:

— differences in smoothing and HRF model

— differences in smoothing and motion regressors

— differences in motion regressors and HRF model

For the first set of between-group analyses (5mm, no derivatives > 8mm, presence of derivatives), under SPM, the results were close to those obtained for the analyses “5mm > 8mm" (from 0.020 to 0.081 in the first case and to 0.101 in the second, for the analysis using pipelines with

24 motion regressors each) (Fig. 6.6). The effect of the HRF derivatives was not very important, as it was for the isolated analyses “no derivatives > presence of derivatives". Similarly, under FSL, these previous analyses gave slight conservativeness, and adding it to the difference in smoothing here gave a slightly lower increase in false positive rate (from 0.014 to 0.058) than for "5mm > 8mm" alone (from 0.014 to 0.082). Similar observations can be made on the P-P plots on Figure 6.8 and 6.11. The absence of noticeable effect of the HRF derivatives can also be seen when combined with differences in numbers of motion regressors in SPM; in FSL, the HRF derivatives have a stronger negative effect of the statistical values and the number of motion regressors a lower positive effect, which makes them compensate each other.

For the second set of analyses (5mm, 24 motion regressors > 8mm, no motion regressors), we can observe on Figures 6.9 and 6.12 that for both software packages, the distribution obtained combined the effects on mean and variance that we observed for both comparisons with a single difference. Thus, as each isolated difference gave invalid results and an inflated false positive rate, their combination gave even more important invalidity, as we can observe on Figure 6.11 and 6.8, and one of the highest observed false positive rates for our study (going from 0.020 to 0.194 in SPM, and from 0.014 to 0.208 in FSL) (Fig. 6.6).

6.4 Discussions

We observed that, when combining data with different pipelines, the effect of differences in subject-level processing depended on which parameter differed between the pipelines. Some differences in parameters gave conservative results, same as what we observed for identical pipelines analyses, while other differences gave invalid results. For this reason, there are no guarantees regarding the validity of positive results that could be obtained, in practice, when performing between-group analyses under similar circumstances, using subject data processed differently. If significant differences are observed in experiments under these conditions, it cannot be determined whether these differences are associated to an existing effect or to the differences in processing.

Our results showed that when performing analyses using the same subject-level pipeline on all subject data (as is traditionally done in the existing literature), results were valid for all analyses. For this reason, the invalid results obtained when combining subject data processed differently tell us that it is necessary to consider how analytical variability may affect the results when combining data, and not that one or the other subject-level pipeline is responsible for the invalidity of the results and should not be used in any case. Although the false positive rates obtained in this situation were lower than the expected 5% rate, the results were similar to those obtained for a similar framework in [40].

Our results for different pipeline analyses suggest that some combinations of pipelines should

be avoided, while others could be used. We saw that for differences regarding the presence of temporal derivatives of the HRF, the results were similar to those obtained with identical pipeline analyses, suggesting that subject data can be combined without having to consider the differences in pipelines, if this is the only difference. This is not the case for differences in smoothing and number of motion regressors, which gave invalid results. For motion regressors, the amplitude of the effect depended on the parameter values chosen in each pipeline. We also saw that combining multiple differences in parameters could result in bigger effects. Our observations of the distributions of statistical values also allowed us to have an insight at how each difference in parameter affected the results (differences in mean and variance compared to the expected distribution), as well as their combination.

For each variation of parameter between pipelines, we saw consistent effects, across the two software packages under study (SPM and FSL), and with different values for the non-varying parameters: for example, all analyses using pipelines with the same number of motion regressors, the same model of the HRF and different levels of smoothing (5mm and 8mm) gave a similar inflation of false positive rates, for all possible cases of numbers of motion regressors and models of the HRF (identical in both pipelines). This, and the effects observed with combined variations of parameters, suggests that it may be possible to model the effect caused by specific variations in the subject-level pipelines. To do this modelisation with processed subject data taken from existing datasets, the pipelines used on the data have to be shared in a format that would allow to know and reproduce the exact processing applied on the data.

While our study shows that invalid results can be obtained when performing between-group analyses where each subject data is processed with a different subject-level pipeline depending on the group, one may wonder how common it would be for such a situation to occur in practice. As said in the introduction, this is not the only type of situation which may appear where there are combinations of subjects data processed differently: subject data from multiple subject-level pipelines may be combined within a group, and multiple groups may contain data processed with a same subject-level pipeline.

The situation that we have here, where subject-level pipelines differ depending on the group, may occur for example when using data from various datasets, if the subject data within each group comes from the same dataset. For example, specific datasets have been created to study various neurological disorders (Alzheimer's Disease Neuroimaging Initiative (ADNI) [79] for Alzheimer's disease, Autism Brain Imaging Data Exchange (ABIDE) [32] for autism, etc), and researchers may want to use them to compare groups of subjects where each group corresponds to a specific disease. If the datasets that they want to use for these comparisons only include processed subject data, it is likely that, within each dataset, processing applied on the subject data will be the same, and that across datasets, processing will differ. Each dataset may then be associated to a specific processing pipeline for its subject data and consequently, each group

also. In this situation, in practice, using the same pipelines that we used here will have an effect which is similar to the one observed here. Also, since the methodological choices made to create the pipelines that we used (in terms of steps performed and chosen parameter values) are very common in the literature [20], the resulting pipelines are typical examples of pipelines used in task-fMRI.

We chose to study variations induced by 3 types of parameters (HRF, smoothing, motion regressors), within each software package, but there are many more used in practical conditions: researchers might use different software versions, perform or not specific substeps within the analysis (for example, slice-timing correction), using different models of the HRF, etc. In real conditions, if researchers combine subject data processed differently, the differences between pipelines will likely be more important.

In future works, other analyses may be done for other varying parameters such as the ones mentioned above, with the same framework that we used in this study. Doing so would provide a more precise knowledge of the real extent of the impact of analytical variability, and maybe lead to observations of phenomenons that are different from what we observed here. For example, even though our study show no evidence that this is the case, differences in subject-level pipelines with an effect which would be the opposite of what we have seen here (lower detection than with identical pipelines) may exist: in this case, the differences in subject-level pipelines could also make it harder to detect existing effects. Also, using a similar framework with paradigms other than the motor task would give us a more precise idea of each parameter's effect in different situations. For example, the presence or absence of HRF derivatives may have a more important impact when using a paradigm with event-related design than in our situation.

An example of potential use case where this framework may be applied is the case of datasets using adaptable pipelines to apply processing steps on the subject data depending on which data is available. This may help to decide whether, for group analysis, it would be relevant to use the processing steps of the adaptable pipelines on the available data, which may differ from the ones applied on other data, rather than using processing steps that match those of other data to have consistent processing across all subject data.

We have seen that the effects of analytical variability often prevent us from combining data without considering the differences in processing pipelines. Our observations on the data suggest that this effect may be corrected, and thus it may be possible to overcome the problem of analytical variability when combining data. For other sources of variability, methods have been proposed to correct this effect: for example, correcting the variability resulting from imaging site and scanner effect (technical variability) in neuroimaging [8, 47]. Finding solutions for issues related to analytical variability in general is a growing research subject in many fields: for example, finding a consensus between different results obtained with different methods, which would correspond to results obtained with different complete processing pipelines [30]). Our

study has already allowed us to measure the impact of analytical variability on the combination of data processed differently for a set of specific varying parameters, and gives a framework for future studies to measure it for other parameters. The ability to combine processed data easily would be a major step toward more reproducible research, and the correction of the effect of analytical variability in such a situation is still a widely unanswered research question. Therefore, investigating ways to solve this issue may become a topic of increasing interest in the near future for the neuroimaging community.

6.4.1 Conclusion

Our study shows that, when combining subject data which have been processed differently, the validity of the results obtained depend on the differences between the pipelines used on subject data. While there are parameters for which differences in values between pipelines do not seem to have major effects, such as the presence of temporal derivatives of the HRF, some other parameter differences can produce invalid results, suggesting that it is impossible to combine processed fMRI data without taking into account these differences in subject-level processing. In further works, we will create methods to correct the effect of analytical variability that we observed in this situation.

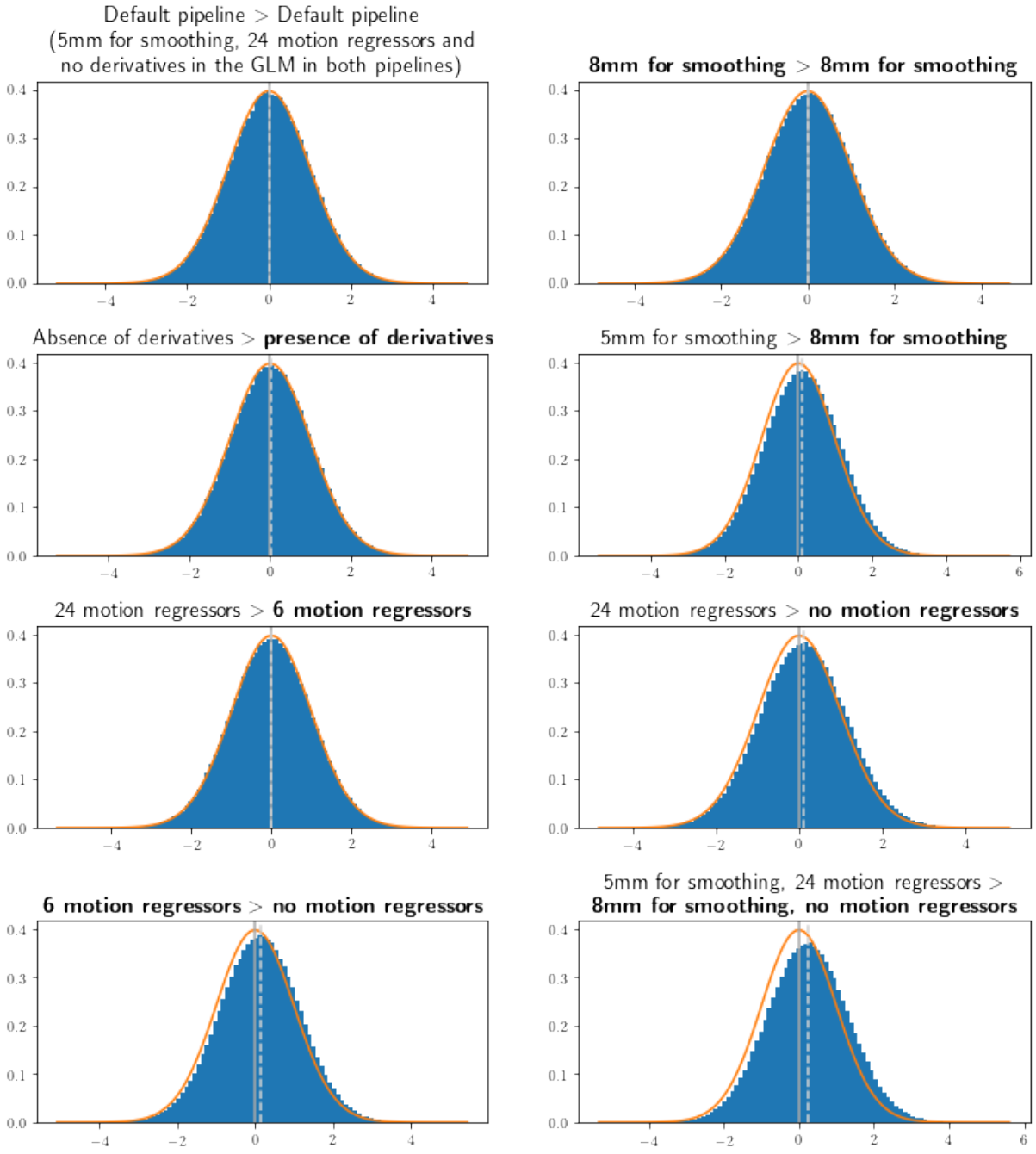


Figure 6.9 – Distribution of statistical values for multiple between-group analyses under SPM, compared to the expected distribution. Pipelines are defined by their differences with the default parameters (5mm for the smoothing kernel FWHM, 24 motion regressors and no temporal derivatives of the HRF in the GLM). Pipelines which differ from the default pipeline are put in bold. The orange curve represents the Student distribution with 98 degrees of freedom, which is the expected distribution in our case under null hypothesis. Means of the obtained distribution are indicated by the dotted grey line.

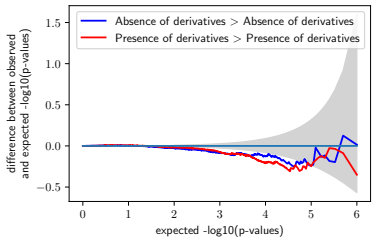
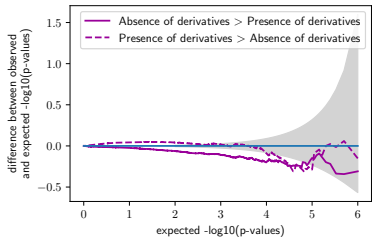
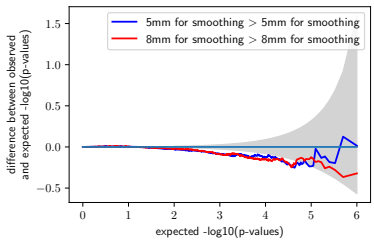
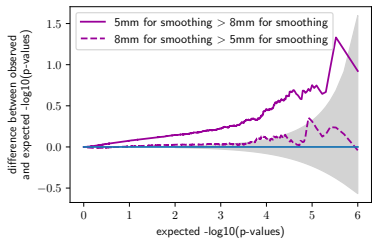
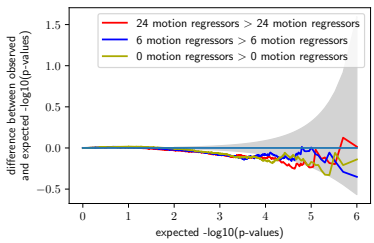
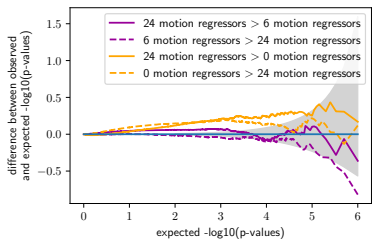
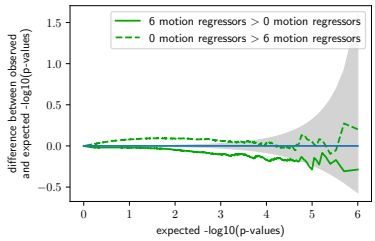
	Identical Pipelines	Different Pipelines
Effect of HRF		
Effect of Smoothing		
Effect of Motion Regressors		 

Figure 6.10 – variants of P-P plots for distributions obtained with various analyses under FSL, against the expected distribution, with 0.95 confidence interval, with a single varying parameter for different pipelines analyses. The variations from usual P-P plots are the use of -log(p-values) instead of p-values (to have a more precise observation of the behavior in the right tail) and replacing the obtained -log(p-values) on the y-axis by the difference in obtained and expected -log(p-values).
When not indicated, the default parameters for the pipelines were: 5mm for the smoothing kernel FWHM, 24 motion regressors and no temporal derivatives of the HRF in the GLM.
Positive differences indicate invalidity whereas negative differences indicate conservativeness.

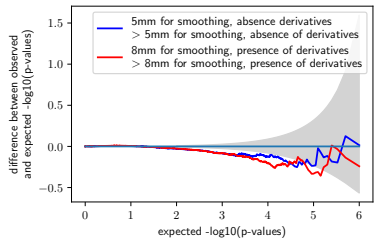
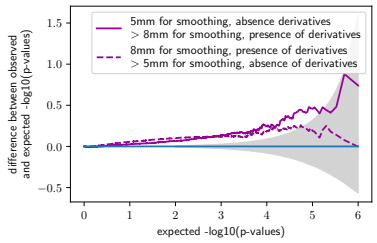
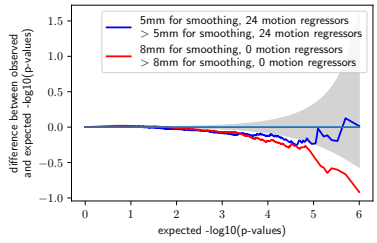
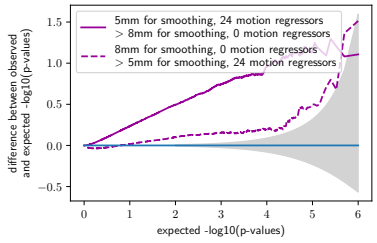
	Identical Pipelines	Different Pipelines
Combined Effect of Smoothing and HRF		
Combined Effect of Smoothing and Motion Regressors		

Figure 6.11 – variants of P-P plots for distributions obtained with various analyses under FSL, against the expected distribution, with 0.95 confidence interval, with a combination of varying parameters for different pipelines analyses. The variations from usual P-P plots are the use of -log(p-values) instead of p-values (to have a more precise observation of the behavior in the right tail) and replacing the obtained -log(p-values) on the y-axis by the difference in obtained and expected -log(p-values).
When not indicated, the default parameters for the pipelines were: 5mm for the smoothing kernel FWHM, 24 motion regressors and no temporal derivatives of the HRF in the GLM.
Positive differences indicate invalidity whereas negative differences indicate conservativeness.

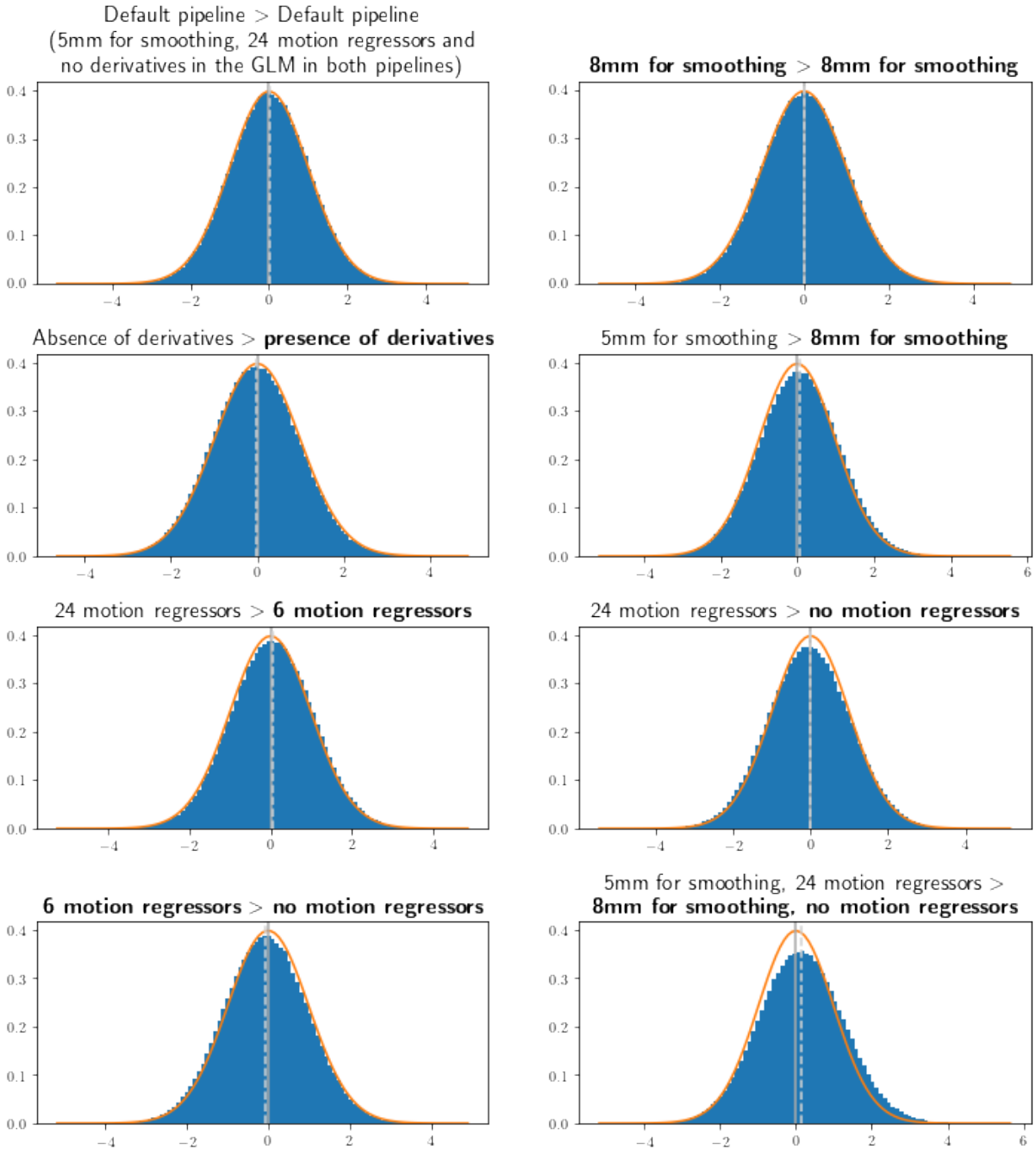


Figure 6.12 – Distribution of statistical values in between-group analyses for multiple pairs of subject-level pipelines under FSL, compared to the expected distribution. Pipelines are defined by their differences with the default parameters (5mm for the smoothing kernel FWHM, 24 motion regressors and no temporal derivatives of the HRF in the GLM). Pipelines which differ from the default pipeline are put in bold. The orange curve represents the Student distribution with 98 degrees of freedom, which is the expected distribution in our case under null hypothesis. Means of the obtained distribution are indicated by the dotted grey line.

Chapter 7

Second Contribution: Correction of Pipeline Effects in Group Analyses

Our work in chapter 6 showed that combination of data processed differently can lead to an increased probability of finding false positive results. The false positive rates obtained when combining data processed differently depended on the parameters of the pipelines that were changed. Some pipeline differences had little to no effect on the results (such as the absence or presence of temporal derivatives of the HRF in the GLM model at first-level). On the other hand, differences in level of smoothing and number of motion regressors yielded inflated false positive rates. This reduces the interest of statistical power achieved by combining data.

However, data combination can still be an interesting solution to the problem of low sample sizes if methods can be applied to assess and correct the effect of processing differences in pipelines. The question of how the effects of variability can be corrected is a common problem in research, in different scientific fields and for multiple sources of variability. In this chapter, we will present a review of the literature on different existing methods for the correction of variability, and their fields of application. We will also present preliminary results obtained for a method that we applied to correct pipeline effect, where this effect is estimated within the analyses performed. Finally, we will discuss the potential further works to correct the effect analytical variability on results with combined data.

7.1 Overview of The Literature

The problem of correcting the effects related to sources of unwanted variance, or heterogeneity, in data during analyses is common in various scientific fields. This includes neuroimaging, as we have seen, with the different sources of variability in the data (including analytical variability), but also fields such as genomics [82, 55, 94], which uses microarrays to detect expression

level for a large number of genes, resulting in high-dimensional microarray data. Examples of sources of unwanted variations in these data include non-biological effects, which are linked to the conditions of acquisition of the data. Methods have been developed to correct those effects, called batch effects, for the study of microarray data. These methods have been adapted for the correction of sources of variability in different neuroimaging modalities [47].

The correction methods can differ depending on the situation in which they can be used, as well as the assumption they make on the data. There are two main situations regarding the approaches that can be used: the case where the variables associated to the source of variation are known in the data, and can be directly modeled within the analysis, and the case where they are unknown (or there are multiple sources which are too complex to model directly) and have to be estimated from the data. In this section, we will give an overview of some of the most popular methods for correction of variability that have been developed in genomics and how they have been adapted for neuroimaging.

7.1.1 ComBat

Various effects exist in microarray data: effects related to biological factors, but also effects related to the differences resulting from combining data sampled at different times, under different conditions. Factors for these technical effects, called batch effects, include conditions regarding the material of data acquisition, atmosphere, temperature, and time of experiments. Methods have been developed to correct these effects in microarray data studies [135, 150].

A simple correction method is the use of Location and Scale (L/S) adjustments. The L/S model assumes that there are additive and multiplicative parameters, which depend on genes and batches, applied over an error term for each gene expression value. An expression value for a gene g in a sample j from a batch i is given by the following formula:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg}$$

where α_g is the gene expression which depends only on the gene g, X a design matrix for sample conditions and β_g a vector of regression coefficients corresponding to X. The batch effects, depending on batches and genes, are given by the additive parameter γ_{ig} and the multiplicative parameter δ_{ig} over an error term ϵ_{ijg} following a Normal distribution with expected value equal to 0 and variance equal to ${\sigma_g}^2$.

The goal of L/S adjustments is to build a batch-adjusted value Y^*_{ijg} which corresponds to what Y_{ijg} would be if there were no multiplicative and additive parameters applied on the error term ϵ_{ijg}. Since α_g and β_g are the same for a fixed gene g, and γ_{ig} and δ_{ig} are the same for fixed batch i and gene g, the observed values Y_{ijg} can be used to build estimates estimates $\hat{\alpha_g}$, $\hat{\beta_g}$, $\hat{\gamma_{ig}}$ and $\hat{\delta_{ig}}$. These estimates are then used to build the batch-adjusted value:

$$Y^*_{ijg} = \frac{Y_{ijg} - \hat{\alpha_g} - X\hat{\beta}_g - \hat{\gamma_{ig}}}{\hat{\delta_{ig}}} + \alpha_g + X\hat{\beta}_g$$

ComBat [82] is an empirical Bayes method which uses the same model for expression value for a gene for given batches and samples, but makes specific assumptions regarding the similarities in the way batch effects affect the genes. The aim of ComBat is to improve the estimates obtained when batch sizes are low. The first step is to perform gene wise standardization, using the following standardized data:

$$Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha_g} - X\hat{\beta}g}{\hat{\sigma_g}}$$

Unlike L/S, Combat does not uses an estimate of the multiplicative parameter $\hat{\delta_{ig}}$, but an estimate of the variance $\hat{\sigma_g}$ (which is built using the estimates $\hat{\alpha_g}$, $\hat{\beta}_g$, $\hat{\gamma_{ig}}$). ComBat then assumes that the values Z_{ijg} for a same batch i and a same gene g follows a normal distribution $N(\gamma'_{ig}, \delta^2_{ig})$. Using either parametric prior regarding the distributions of γ'_{ig} (Normal distribution) and δ_{ig} (Inverse Gamma distribution), or non-parametric prior, adjusted batch effect estimators γ^*_{ig} and δ^*_{ig} are built to calculate the adjusted γ^*_{ijg}.

7.1.2 Surrogate Variable Analysis

Methods such as ComBat require to have information regarding components of expression heterogeneity. However, this information is not always available, or may be too complex to be directly modeled precisely. For this reasons, other methods can be used to identify and estimate these components. This is the case of Surrogate Variable Analysis (SVA) [94], which builds these components from the data.

In the case of genomics, we consider a matrix representing m genes for n arrays. the SVA procedures goes as follows: first, a residual expression matrix is built by removing the signal related to the primary variable of interests. A decomposition is applied on this matrix to create singular vectors that are associated to signatures of expression heterogeneity, and the most important are selected using a statistical test to determine whether they represent a significant part of variation in the data. Subsets of genes corresponding to each signature are then determined using another statistical test, and these subsets of genes are used to build surrogate variables which are included in the model.

SVA can be useful when there are latent sources of variations that may not be incorporated directly and need to be identified beforehand if researchers have to correct them. However, these latent variables may also be correlated with variables of interest. Therefore, it can be difficult to determine whether or not it may improve the inference. [93]

7.1.3 Remove Unwanted Variation

Similarly to SVA, the family of “Remove Unwanted Variation" (RUV) methods [55], also designed primarily for the correction of batch effects in microarray data, uses a correction on the data without prior knowledge about the effect. The main idea behind RUV is to take advantage of the presence of negative controls, which are assumed not to be affected by the variables of interest, to estimate the unwanted variation that we want to correct. For microarray data, let us consider a matrix $Y_{m\times n}$ giving the log expression of n genes on m arrays. The model for RUV is the following:

$$Y_{m\times n} = X_{m\times 1}\beta_{1\times n} + W_{m\times k}\beta_{k\times n} + \epsilon_{m\times n}$$

where X is a column matrix corresponding to a single factor of interest, β the set of unobserved parameters that we want to estimate for each gene, W a matrix of unobserved variables (the number of unobserved variables k has to be defined before analysis), α a set of either random or fixed parameters and ϵ_{ij} error terms which are independent and have a same normal distribution with expected value of 0 and variance ${\sigma_j}^2$.

The idea of RUV is to use a subset Y_c of Y consisting of control genes, such that the associated β_c are supposed to be equal to 0. Here we will present some of the variants of RUV in chronological order of appearance in the scientific literature.

RUV-2

The most simple RUV method is the 2-step method (RUV-2) [55]. We have the following formula:

$$Y_c = X\beta_c + W\alpha_c + \epsilon_c$$

Using the assumption $\beta_c = 0$, the formula becomes:

$$Y_c = W\alpha_c + \epsilon_c$$

The first step is to apply factor analysis on Y_c to estimate W. The second step is to estimate β with a regression of Y on X and the estimate of W.

With $R_A = I - A(A'A)^{-1}A'$ being the residual operator of a matrix A (orthogonal projection on the column space of A), the estimate $\hat{\beta}$ is given by the following formula:

$$\hat{\beta} = (X'R_{\hat{W}}X)^{-1}X'R_{\hat{W}}Y$$

RUV-4

RUV-4 [54] is a version of RUV based on both RUV-2 and SVA. The main issue regarding the fact that unwanted factors are unknown in RUV and SVA is that signal of interest may be removed along with unwanted variance. Instead of solving this problem directly like RUV-2, by performing factor analysis on W using only the control genes, RUV-4 does it by performing factor analysis on the residual component $W_0 = R_X W$. The graphical description of estimates used in RUV-4 is shown on Figure 7.1. The first two steps are the following: the X component is removed from the equation using R_X : $R_X Y = R_X W\alpha + R_X\epsilon$. A factor analysis is done to produce an estimate $\hat{W_0\alpha}$ of $W_0\alpha$, with $W_0 = R_X W$.

The problem then consists in estimating W from the estimate of W_0, before performing regression as in RUV-2. If we define b_{WX} as the partial regression coefficient of W on X ($b_{WX} = (X'X)^{-1}X'W$), we have $W = W_0 + Xb_{XW}$. The goal of RUV-4 is to use the control genes like RUV-2 to estimate b_{XW} at this point. Using a decomposition of Y with a control gene component Y_c, and the approximation $\alpha_c \approx \hat{\alpha_c}$, we find the following estimation:

$$\hat{W} \approx \hat{W_0} + Xb_{Y_c X}\hat{\alpha_c'}(\hat{\alpha_c}\hat{\alpha_c'})^{-1}$$

Finally, we apply a regression to estimate β using our estimates $\hat{W}$ as we did for RUV-2:

$$\hat{\beta} = (X'R_{\hat{W}}X)^{-1}X'R_{\hat{W}}Y$$

7.1.4 Applications to Neuroimaging

We have seen in chapter 5 that, for various neuroimaging modalities (including task-based BOLD fMRI), there exist multiple sources of variability which may impact results and that researchers may want to correct when doing studies. For this reason, methods developed for genomics such as those detailed here have been adapted for neuroimaging. Usually, the idea is to apply correction on the voxel values in brain imaging data.

An example of adaptation for neuroimaging is Removal of Artificial Voxel Effect by Linear regression (RAVEL)[47], which is a method inspired by RUV for the correction of technical variability, with unwanted variation component estimated from the cerebro-spinal fluid. A method such as ComBat has also been adapted in the case of neuroimaging, with voxels values replacing genes log expression, and modified for longitudinal studies also [8]. It has also been used for measurements of cortical thickness for example [45], or for diffusion tensor imaging, along with RAVEL and SVA [46].

In neuroimaging, these methods mostly address the question of what we defined in chapter 5 as technical variability, which is related to scanner and site effects. They are all used in a situation

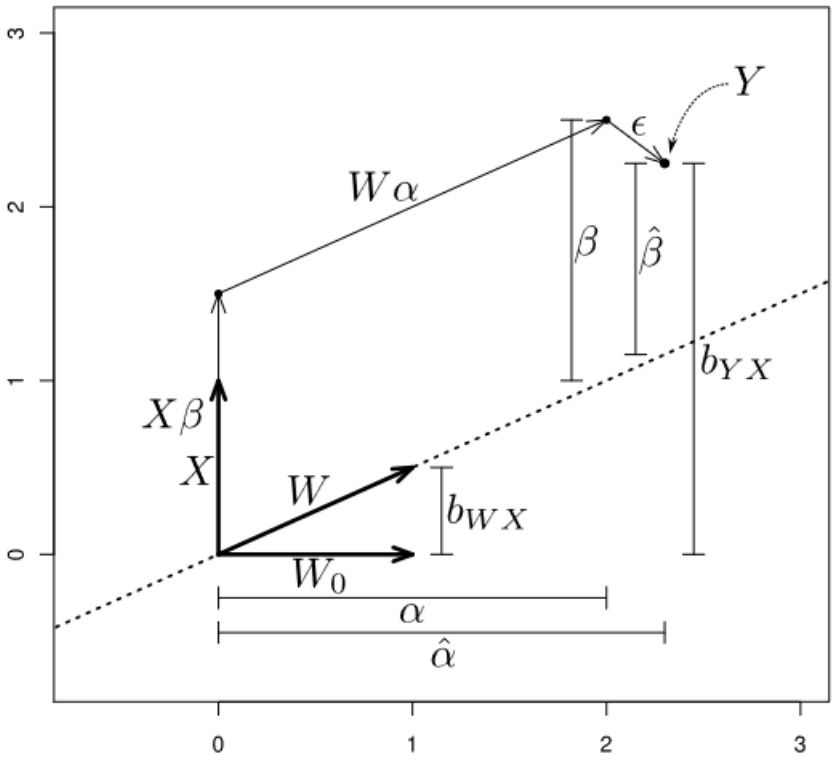


Figure 7.1 – Graphical depiction of the method for RUV-4, from [54]. The method uses an estimate of the projection $W_0\alpha$ of $W\alpha$ as an intermediate step to estimate W before estimating β through regression.

that bears similarities to ours, with a necessity to combine data to take advantage of large sample sizes, and may be used for the case of analytical variability, depending on the situation where we want to apply correction. As there are many steps in fMRI processing, correction of analytical variability may be applied at different levels.

7.2 Methods

In our work, we have seen in chapter 6 that analytical variability has an effect on fMRI study results when combining data processed differently, and we may want to use methods to estimate and correct this effect. Multiple options are possible regarding the correction of effects of analytical variability in between-group analyses combining data processed differently, depending on the situation that we have. In particular, adaptations of the various methods presented in the previous section, used in different fields to correct the effects of specific sources of variability, can be developed to remove the variance related to differences in subject-level processing in fMRI.

However, the adaptation of these methods from their original field of application to neuroimaging can lead to non-trivial questions regarding what components in the data take the different roles in the model.

For this reason, we chose to begin by evaluating a method which is more directly adapted for the situation that we have here, where we use a GLM for statistical analysis. The method that we use here consists in adding an extra binary covariate in the second-level analysis design matrix, where values associated to each subject are equal to 1 or 0 depending on which pipeline was used to process subject data, in order to estimate and correct the effect of analytical variability. This method is a simple way of adding a component corresponding to the effect of variability in the modeling of the signal of interest, as it is often done in group studies in fMRI [127]. In the initial second-level GLM, the estimated value of the first-level contrast is modeled as the sum of the mean value within the subject's group, and random noise. In the new version, it is modeled as the sum of the mean value within the subject's group, the effect of pipeline difference, and random noise.

Compared to more complex methods, this method has the advantage of allowing the estimation and correction of the pipeline effect directly within the analysis. The only necessary information is knowing which subjects in each group have been processed by each pipeline. This method does not require to do other analyses independently for the estimation, nor does it require to have information about the nature of the differences between pipelines. Besides this, this approach is also easier to implement than other models which are more sophisticated.

In our previous study, we performed analyses using two subject-level pipelines where each group data was entirely processed by one of both pipelines. However, our method cannot be used if there is colinearity between covariates in the second-level design matrix. This means that there cannot be a group with all subjects processed by one pipeline, and a group with all subjects processed by another, as done previously. Otherwise, the between-group difference that we want to estimate will be confounded with the between-pipeline difference that we want to remove. Because of this, our method of correction can only be applied in between-group analyses where each group contain subject data processed by both pipelines.

For this reason, the between-group analyses performed were done using the pairs of pipeline defined previously with specific proportions of subjects processed with each pipeline in each group, as described on Figure 7.2. The first group contained a proportion p_1 of subject data processed with the first pipeline and a proportion $1 - p_1$ with the second pipeline, and the opposite for the second group. Analyses were performed with proportions p_1 equal to 50% (strict equality between both groups), 70%, 80% and 90%, with and without correction applied. Results without a covariate for correction can be compared to the results obtained in the previous study to observe what is the impact of the proportion on the effect of pipeline differences. Results with a covariate for correction can be compared to the results without covariate to see what is the

impact of the correction method at different proportions.

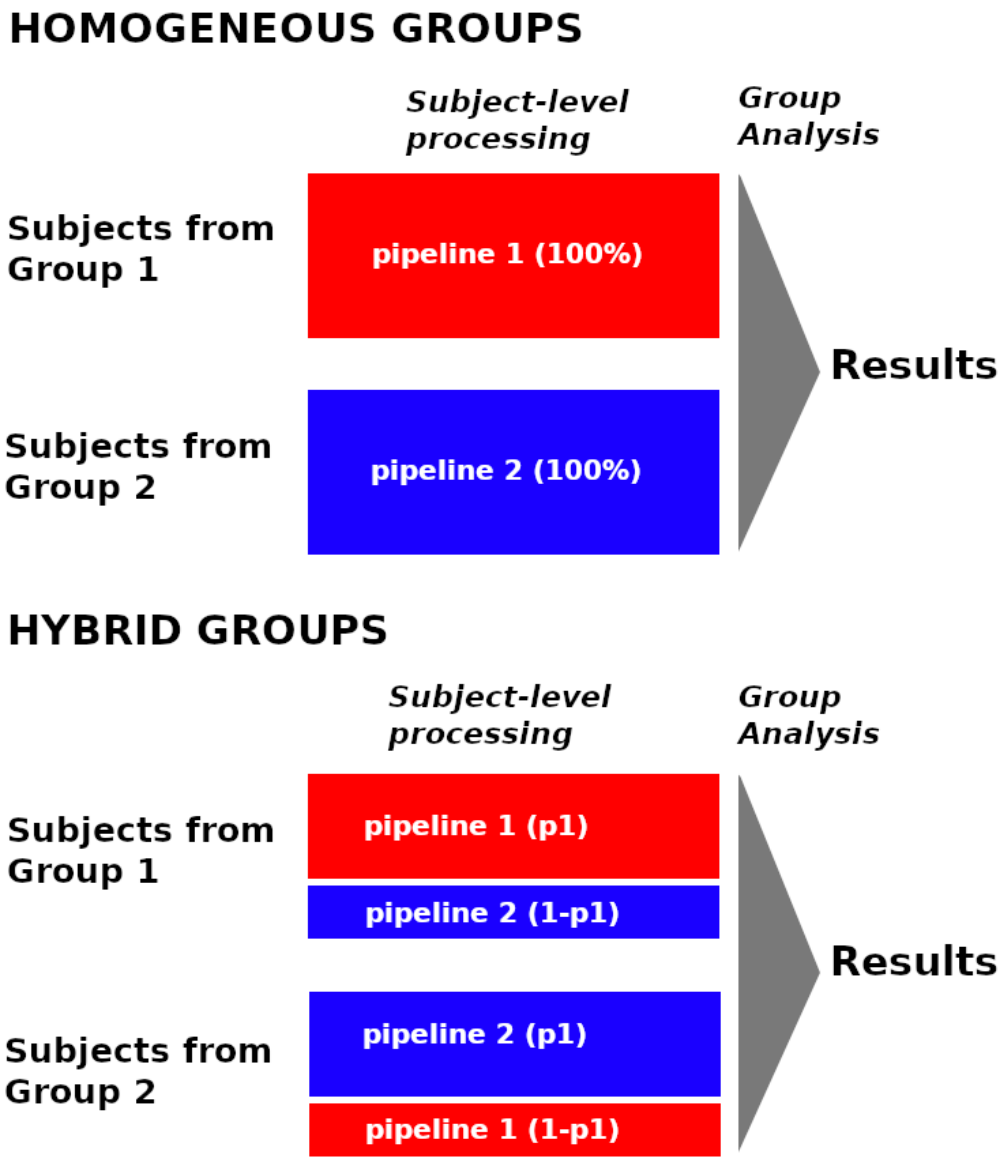


Figure 7.2 – Difference between analyses performed in our work in chapter 6 and in this work. In our previous work, between-group analyses were performed with all data within a group processed using a same pipeline. In our present work, we perform group analyses where each group contains a mix of subjects processed with both pipelines used, with proportions p_1 / $1-p_1$ of subjects with data processed with each pipeline for the first group and $1-p_1$ / p_1 for the second group. Analyses are done for the following values for the proportion p_1: 50% (equal proportions for both pipelines), 70%, 80% and 90%.

To assess the performance of the correction method applied, we estimated the validity of

analyses with and without covariates for correction using the same framework that we used in our previous work. We performed between-group analyses, in which we observed the difference in mean activation between two groups of 50 subjects, for a first-level contrast corresponding to the right hand, using data from subjects in the HCP dataset who had completed the motor task. This analysis was repeated 1,000 times with different pairs of groups (the set of 1,000 pairs of groups was the same that was used previously).

As in chapter 6, we computed the p-values associated to a null hypothesis of no difference of mean activation between both groups. Since the pairs of groups are constructed so as to verify the null hypothesis, we expect the p-values to be uniformly distributed on $[0, 1]$. We compared the distributions of obtained and expected ordered p-values using variants of P-P plots, as we did in the previous chapter.

Results were obtained for analyses using subject data processed by two different pipelines, for various pairs of pipelines. These pairs of pipelines are defined as follows: a first pipeline always equal to the default pipeline defined in the previous study, and a second pipeline which varies from the default pipeline on one parameter (smoothing, motion regressors, temporal derivatives of the HRF) or two (smoothing and motion regressors combined). Subject-level and group-level processing of the data was performed using SPM12 r7771 (RRID: SCR_007037, [118]) with Octave 5.1.0 (RRID: SCR_014398, [38]).

7.3 Results

7.3.1 Results Without Correction

Figures 7.3 and 7.4 shows P-P plots obtained for analyses with the proportions that we defined, for all four pairs of pipeline, both without correction (in blue) and with correction (in red). At 50%, the P-P plots obtained show no invalidity, and are similar to the results obtained in our previous work for identical pipeline analysis (One same pipeline applied on all subjects in both groups).

With higher proportions, we observe results which tend to become similar to the ones observed with 100% of subjects within a group processed by a single pipeline (as we had in our work in chapter 6). There is no major effect of the HRF, effect of smoothing and regressors and combination of parameter differences (smoothing and regressors) gives a stronger effect than each of them isolated. The amplitude of the effect tend to increase with the proportion.

7.3.2 Results With Correction

We have seen that the presence of effects depends on the proportion of subjects processed with each pipeline within each group: the effect that we saw in our previous works are absent at 50%, and tend to appear as we draw closer to 100%. Figures 7.3 and 7.4 also shows the P-P

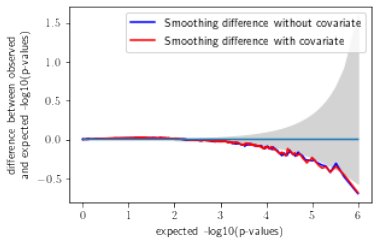
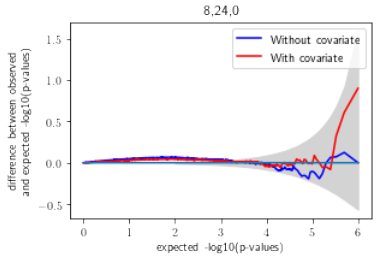
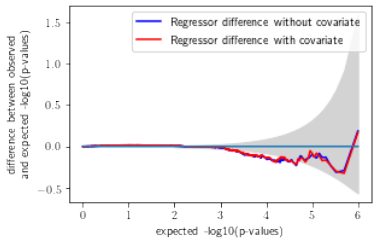
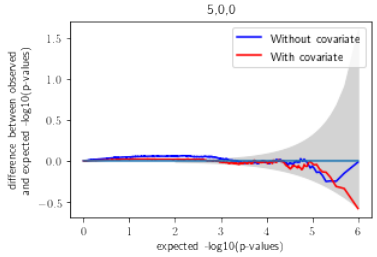
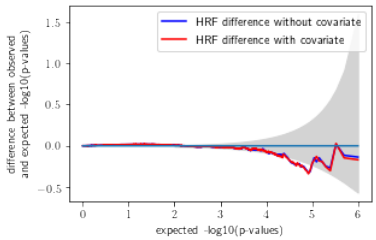
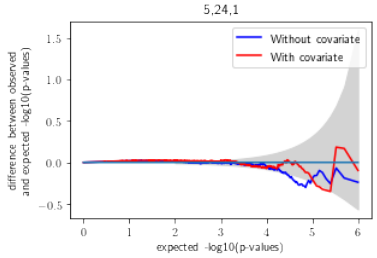
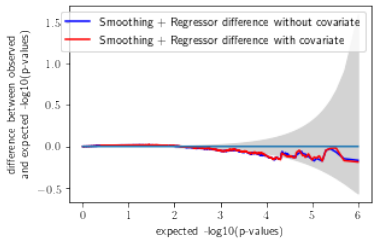
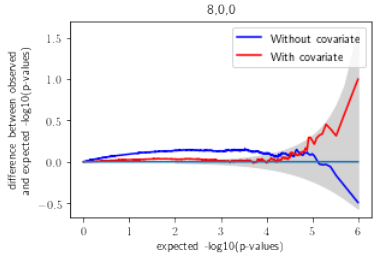
Pipeline proportions	50% / 50%	70% / 30%
Effect of Smoothing		
Effect of Regressors		
Effect of HRF		
Effect of Motion Regressors + Smoothing		

Figure 7.3 – Variants of P-P plots obtained for between-group analyses performed under SPM combining data processed with different pipelines, both with and without a covariate for correction. For each analysis, the first pipeline was the default pipeline defined in chapter 6, and the second pipeline was a pipeline differing from the default pipeline on one or two parameter values. Results are shown for analyses with the following proportions $p_1/1-p_1$ of subjects processed with each pipeline in each group as described in Figure 7.2: 50%/50% for analyses in the first column, 70%/30% in the second column.

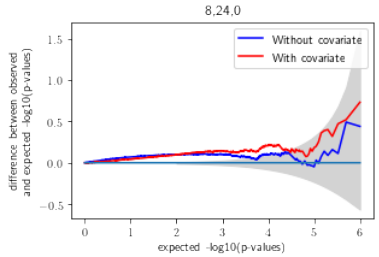
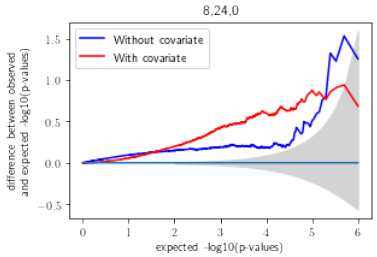
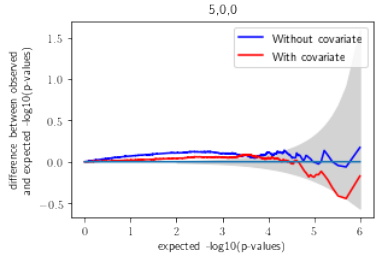
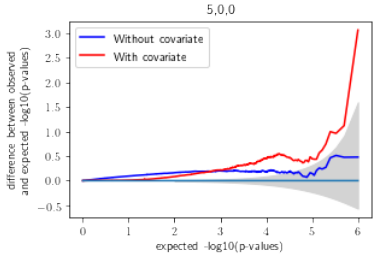
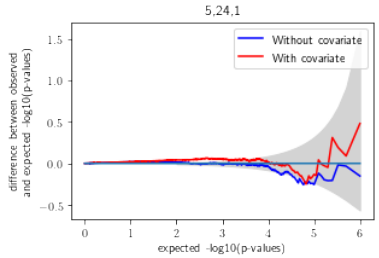
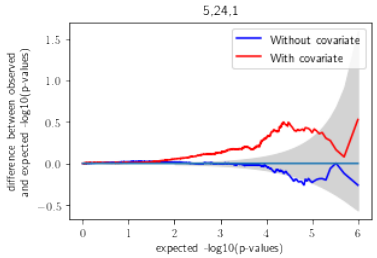
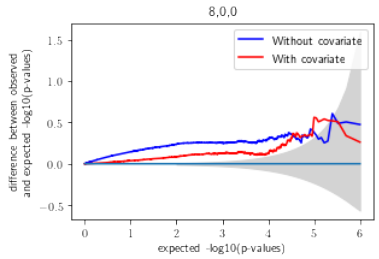
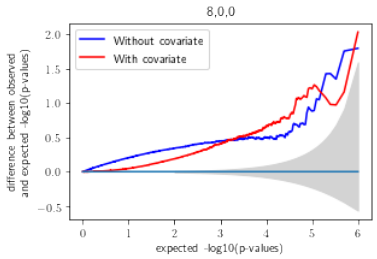
Pipeline proportions	80% / 20%	90% / 10%
Effect of Smoothing		
Effect of Regressors		
Effect of HRF		
Effect of Motion Regressors + Smoothing		

Figure 7.4 – Variants of P-P plots obtained for between-group analyses performed under SPM combining data processed with different pipelines, both with and without a covariate for correction. For each analysis, the first pipeline was the default pipeline defined in chapter 6, and the second pipeline was a pipeline differing from the default pipeline on one or two parameter values. Results are shown for analyses with the following proportions $p_1/1-p_1$ of subjects processed with each pipeline in each group as described in Figure 7.2: 80%/20% for analyses in the first column, 90%/10% in the second column.

plots for analyses at these various proportions, but with a covariate to correct the pipeline effect (in red).

Results with and without the covariate can be compared at the different proportions. At 50%, results with correction are similar to those without corrections, in which there was no observable pipeline effect. At 70%, and 80%, where the effect of pipeline differences begin to appear, adding the covariate seems to partially erase this effect, especially for differences in motion regressors and combination of motion regressors and smoothing. At 90%, however, the correction does not seem to work, as the curves for analyses with covariates do not correct the invalidity and even decrease the obtained p-values in the tail, notably in the case of differences in HRF where invalidity appears although there was none beforehand.

7.4 Discussion and Further Works

In our previous work, we defined a default pipeline with a smoothing kernel of 5mm FWHM, 24 motion regressors in the first-level GLM design matrix and no derivatives of the HRF. We notably observed results for between-group analyses where all the subject data in the first group was processed with the default pipeline, and all the data in the second group processed with a pipeline varying from the default pipeline on one or multiple parameters. We also observed in our results that there was no noticeable effect with differences in the presence or absence of derivatives of the HRF. We also saw that a variation in smoothing (smoothing kernel FWHM equal to 8mm) or number of motion regressors (none) in the varying pipeline led to an increase in false positive rates. Finally, combination of parameter differences between pipelines led to combination of effects associated to these differences (for differences in smoothing and regressors for example).

We performed analyses using different proportions of subjects processed by a specific pipeline. Our results showed that when both groups contain the same amount of subjects processed by each pipeline, there is no noticeable effect of the pipeline difference, as we could expect since there are no difference in proportion between both groups. The higher the proportion of a specific pipeline within each group, the closer the effects observed are to those obtained when all subjects are processed with a same pipeline within each group.

We proposed a method to correct the effect of analytical variability when combining data processed differently, simply using a new regressor within second-level statistical analysis corresponding to pipeline effect. The advantage of this method is that it is easy to implement and that the correction applied does not have to be described as a function of the parameter values which differ between pipelines. Also, the effect of pipeline differences is estimated directly within the analysis and does not depend on any other analysis. However, using this method of correction requires to be in a particular situation where the pipeline differences are not confounded with

the group differences.

The efficiency of the method seems to be lower with a proportion p_1 of a specific pipeline within each group equal to 90%. There may be two reasons in particular for this. The first one is the fact that we approach the colinearity that we have at 100% that we wanted to avoid, and the confusion between the difference between groups and the difference between pipelines may have an effect on the estimation of second-level parameters. The second one is the low number of subjects associated to the secondary pipeline in each group when there is a very high proportion of the prominent pipeline. With 50 subjects per group, a 90% proportion of one pipeline means that only 5 subjects are associated to the secondary pipeline in each group. However, this issue can be solved with higher sample sizes.

The framework that we used, where we observed results obtained with and without correction methods in specific analyses, could be used to estimate the performance and relevance of other correction methods. In our literature review, we have presented a variety of methods used in multiple scientific fields to correct the effect of different sources of variability on research results. Such methods may also be adapted in order to be applied in our situation. This could notably lead to research frameworks for the assessment of pipeline effects which differ from ours, for example frameworks where the estimation and correction of pipeline effect is done independently from the analysis where we want to apply the correction. Also, such models may take into consideration what are the specific parameter differences between pipelines and estimate effects and correction methods in function of these parameter differences.

We have tried using correction methods when we combine data processed with two different pipelines, in between-group analysis where each group contain data processed with the two pipelines, in different proportions. In further works, correction methods could be applied in analyses combining data processed with more than two pipelines. Also, with other methods, correction could be applied in the case where each group is associated to a specific pipeline, which was impossible with our method.

We have defined a research framework for the estimation and correction of analytical variability in analyses combining data processed differently, for a given correction method. As for the work done in chapter 6, we may use this framework in other conditions. As we have said, we may try to apply other sorts of correction methods. We may also observe results for different paradigms, and for situations where we have differences across pipelines other than the ones that we studied here. Also, similarly to the work done in chapter 6, our work used between-group analyses with groups constructed to verify the null hypothesis, to be able to estimate the validity of our results. However, it would be interesting to observe the results that we would obtain when performing analyses where we do not put ourselves under the conditions of null hypothesis. This case would likely be a more accurate reflection of the kind of practical situation in which researchers want to address the issue of analytical variability.

Conclusion

The aim of our work was to investigate the impact of different processing pipelines on research results in task-based BOLD fMRI studies when combining subject data processed differently. Performing analyses with subject data all processed identically – as is usually done in fMRI – avoids having to consider the effects of different pipelines. However, being able to combine data processed differently would allow achieving larger sample sizes and help solve the issue of low statistical power which is one of the main obstacle to the reproducibility of experiments in neuroimaging studies.

This raises two key research questions that we addressed in our work. First, we wanted to determine whether or not there was an effect of processing differences which could induce invalidity, and to estimate this effect. Second, we wanted to see, when there was an effect, whether applying certain correction methods would help reduce it in order to obtain valid results. Both of these research questions required building a framework in which we could assess the validity of the results obtained. We know that analytical variability can lead to invalidity even for analyses with identical subject-level processing applied on the subject data. For this reason, simply observing differences in results between analyses with and without a same subject-level processing for all data would have given limited information.

For this reason, we chose to perform analyses which verify the conditions of null hypothesis to take advantage of the theoretical properties that we have in these conditions to have control over our results. Since the groups are built in such a way that we know H_0 is true, we know what results that we are supposed to have (uniform distribution of p-values and 5% false positive rate), but this is only the case in the absence of any pipeline effect.

In our work, we observed the effects of pipeline differences, with or without methods applied to try and correct these potential effects, by performing analyses combining subject data processed with two different pipelines. We did this for multiple pairs of pipelines, with varying values for three parameters: spatial smoothing kernel in preprocessing, number of motion regressors and presence or absence of the derivatives of the HRF in the first-level GLM. Our work was done using data from the Human Connectome Project 1200-subject datasets. For each comparison, we observed the obtained false positive rates and distributions of statistical values, and compared them to the expected results with no pipeline effects.

Our first study consisted in performing between-group analyses, with each group being associated to a specific pipeline. Our results showed that, with the conditions that we chose (contrast associated to the right hand for the motor task, with a block design), differences in smoothing and number of motion regressors had an effect on the mean and variance of the distribution of statistical values, which led to an increase of the false positive rate. On the other hand, no noticeable effect was observed for differences regarding the presence or absence of HRF derivatives. We performed analyses with both subject-level pipelines used under the same software package, either both under SPM or both under FSL. The similarity in results observed with both software packages allowed us to ensure of the reproducibility of our results. This work allowed us to give an answer to our first question: combination of subject data processed differently may increase the chance of obtaining a false positive result, making the analysis invalid. It also allowed us to understand how each specific parameter difference impacted the results.

In our second study, we tried to correct the effect induced by the differences between pipelines by adding a binary covariate in the second-level GLM. This covariate indicates for each subject which pipeline has been used. To estimate the performance of the correction, we performed between-group analyses with and without adding a covariate, with proportions of data processed with each pipeline within each group different from 100%, unlike our previous study. We did this because otherwise, the pipeline difference would be confounded with the group difference, making the design matrix columns linearly dependent. This also allowed us to observe the effect of variability, without correction, in a situation different from the one observed in our first study. For analyses without correction, we saw that the effects observed in our first study were not noticeable when each group contained a same proportion of data processed by each pipeline, and that they appeared when we had a higher proportion of a specific pipeline for each group. When using our correction method, we saw that the effect was partly corrected when the proportions of subjects processed with each pipeline within each group was "reasonable" (70%/30% and 80%/20%). However, at 90%/10%, the correction did not seem to work well. This could be due either to the higher confusion between group difference and pipeline difference, or also to the low number of subjects processed by the secondary pipeline within each group.

There are several ways in which further investigations regarding the effect of analytical variability may be done. Our work was limited to the study of three parameters, but we saw that there is a large number of other parameters which may have an effect on the results. These parameters may also be studied in a framework similar to ours, to observe whether they have an effect on the results at different parameter values. Also, we did a study using the motor task, with a block design, but there are many other possibilities in terms of paradigms. In particular, it would be interesting to observe how the effect of pipeline combination may vary depending on the paradigm (in event-related design compared to block designs, for examples).

We performed analyses where the null hypothesis was true, as it allowed us to establish

a framework in which we could assess the validity of analyses combining data processed with different pipelines. However, it is also possible to perform analyses where we have an actual difference between groups, besides processing differences, and want to observe whether or not this between-group difference has an effect on a given contrast, when combining data processed differently. Doing so would require defining a new framework, with new ways to control the validity of the results (such as defining zones for which no effects of the between-group difference are expected). Other criteria of comparison may be established and other types of observations on the data may be studied.

Further works may also be done regarding the correction of the effects of pipeline differences. As we have seen, the simple addition of a covariate has limitations regarding its efficiency, as well as the situations where it may applied. Other possibilities include correction methods where the pipeline effect and its correction are estimated as functions of the pipeline differences, unlike what we have here; also, methods existing for other sources of variability may be adapted to this situation.

Our work allows us to conclude that, in task-based BOLD fMRI, combining subject data processed by different pipelines in between-group analyses can have an effect on the results. More precisely, there are situations where this effect can make the analysis invalid, i.e. there is an increased chance of detecting an effect when there is none. In this case, the pipeline effect reduces the interest of data combination, which was to increase the probability of detecting a true positive result. We saw that this effect varied depending on multiple factors: what were the pipeline differences, how the data processed with each pipeline were distributed between the two groups. We also saw that, for analyses where we observed invalidity, applying a simple correction method can reduce the pipeline effect observed in some cases.

These results are encouraging regarding the possibility to make analyses combining data processed differently in the future, with further research about pipeline effects and methods to correct them. Finally, the methods and research framework that we developed may be used for other neuroimaging modalities, to the extent where it can be adapted to these situations, as the question of analytical variability is present in fields other than task-based BOLD fMRI.

Appendix

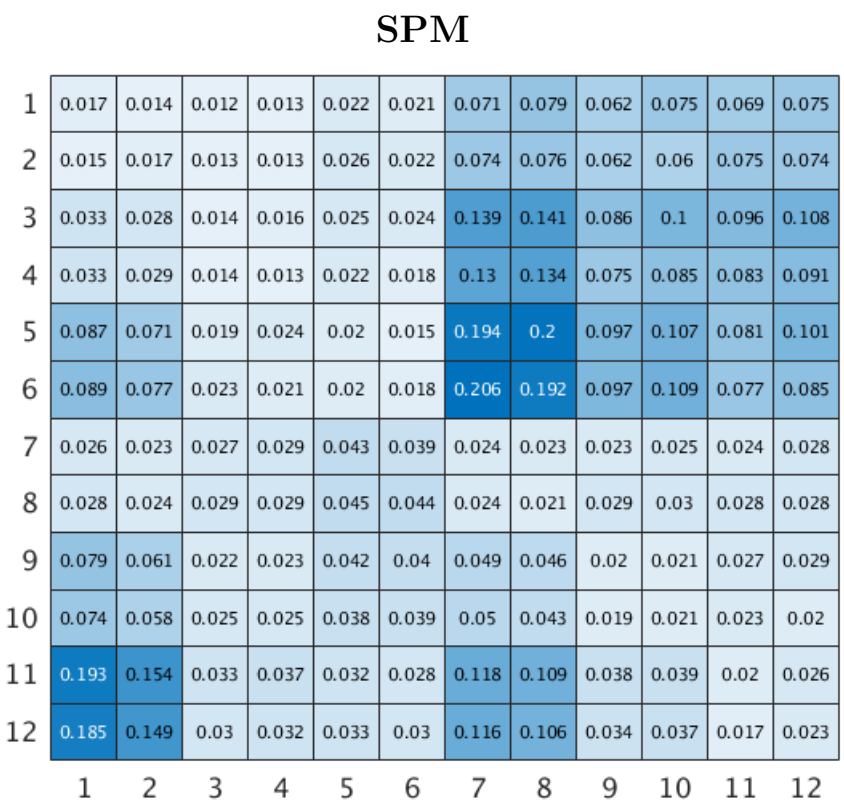


Figure 7.5 – Matrix of false positive rates obtained with SPM for all analyses 'pipeline i > pipeline j' with 1 < i, j < 12. Parameter values for pipeline i are defined as follows, for 1 < i < 12 :
1: 5mm for smoothing, no motion regressors, no derivatives
2: 5mm for smoothing, no motion regressors, presence of derivatives
3: 5mm for smoothing, 6 motion regressors, no derivatives
4: 5mm for smoothing, 6 motion regressors, presence of derivatives
5: 5mm for smoothing, 24 motion regressors, no derivatives
6: 5mm for smoothing, 24 motion regressors, presence of derivatives
7: 8mm for smoothing, no motion regressors, no derivatives
8: 8mm for smoothing, no motion regressors, presence of derivatives
9: 8mm for smoothing, 6 motion regressors, no derivatives
10: 8mm for smoothing, 6 motion regressors, presence of derivatives
11: 8mm for smoothing, 24 motion regressors, no derivatives
12: 8mm for smoothing, 24 motion regressors, presence of derivatives

FSL

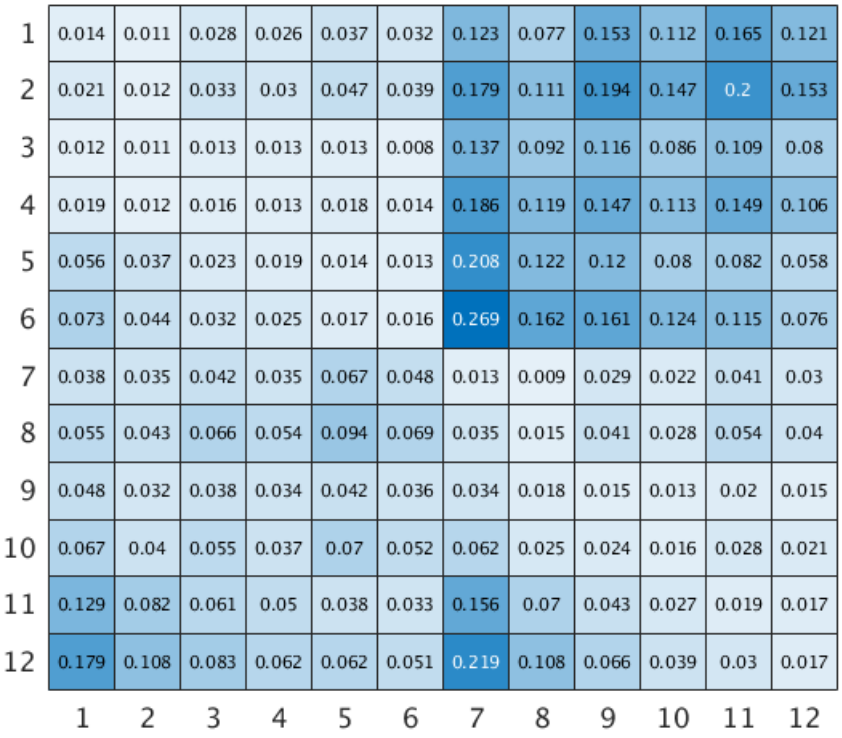


Figure 7.6 – Matrix of false positive rates obtained with FSL for all analyses 'pipeline i > pipeline j' with 1 < i, j < 12. Parameter values for pipeline i are defined as follows, for 1 < i < 12 :
1: 5mm for smoothing, no motion regressors, no derivatives
2: 5mm for smoothing, no motion regressors, presence of derivatives
3: 5mm for smoothing, 6 motion regressors, no derivatives
4: 5mm for smoothing, 6 motion regressors, presence of derivatives
5: 5mm for smoothing, 24 motion regressors, no derivatives
6: 5mm for smoothing, 24 motion regressors, presence of derivatives
7: 8mm for smoothing, no motion regressors, no derivatives
8: 8mm for smoothing, no motion regressors, presence of derivatives
9: 8mm for smoothing, 6 motion regressors, no derivatives
10: 8mm for smoothing, 6 motion regressors, presence of derivatives
11: 8mm for smoothing, 24 motion regressors, no derivatives
12: 8mm for smoothing, 24 motion regressors, presence of derivatives

Bibliography

[1] Ahmed, N., Natarajan, T., and Rao, K. R. (1974). Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93.

[2] Allen, N. E., Sudlow, C., Peakman, T., Collins, R., and Biobank, U. (2014). Uk biobank data: come and get it.

[3] Aron, A. R., Gluck, M. A., and Poldrack, R. A. (2006). Long-term test–retest reliability of functional mri in a classification learning task. *Neuroimage*, 29(3):1000–1006.

[4] Baggerly, K. A. and Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, pages 1309–1334.

[5] Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., et al. (2013). Function in the human connectome: task-fmri and individual differences in behavior. *Neuroimage*, 80:169–189.

[6] Barillot, C., Bannier, E., Commowick, O., Corouge, I., Baire, A., Fakhfakh, I., Guillaumont, J., Yao, Y., and Kain, M. (2016). Shanoir: applying the software as a service distribution model to manage brain imaging research repositories. *Frontiers in ICT*, 3:25.

[7] Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2003). General multilevel linear modeling for group analysis in fmri. *Neuroimage*, 20(2):1052–1063.

[8] Beer, J. C., Tustison, N. J., Cook, P. A., Davatzikos, C., Sheline, Y. I., Shinohara, R. T., Linn, K. A., Initiative, A. D. N., et al. (2020). Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data. *Neuroimage*, 220:117129.

[9] Begley, C. G. and Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533.

[10] Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in science: improving the standard for basic and preclinical research. *Circulation research*, 116(1):116–126.

[11] Biberacher, V., Schmidt, P., Keshavan, A., Boucard, C. C., Righart, R., Sämann, P., Preibisch, C., Fröbel, D., Aly, L., Hemmer, B., et al. (2016). Intra-and interscanner variability of magnetic resonance imaging based volumetry in multiple sclerosis. *Neuroimage*, 142:188–197.

[12] Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., et al. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739.

[13] Blinowska, K. and Durka, P. (2006). Electroencephalography (eeg). *Wiley encyclopedia of biomedical engineering*.

[14] Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88.

[15] Bowring, A., Maumet, C., and Nichols, T. E. (2019). Exploring the impact of analysis software on task fmri results. *Human brain mapping*, 40(11):3362–3384.

[16] Boynton, G. M., Engel, S. A., Glover, G. H., and Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human v1. *Journal of Neuroscience*, 16(13):4207–4221.

[17] Brett, M., Christoff, K., Cusack, R., Lancaster, J., et al. (2001). Using the talairach atlas with the mni template. *Neuroimage*, 13(6):85–85.

[18] Bullmore, E. T., Brammer, M. J., Rabe-Hesketh, S., Curtis, V. A., Morris, R. G., Williams, S. C., Sharma, T., and McGuire, P. K. (1999). Methods for diagnosis and treatment of stimulus-correlated motion in generic brain activation studies using fmri. *Human brain mapping*, 7(1):38–48.

[19] Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, 14(5):365–376.

[20] Carp, J. (2012a). On the plurality of (methodological) worlds: estimating the analytic flexibility of fmri experiments. *Frontiers in neuroscience*, 6:149.

[21] Carp, J. (2012b). The secret lives of experiments: methods reporting in the fmri literature. *Neuroimage*, 63(1):289–300.

[22] Carp, J. (2013). Optimizing the order of operations for movement scrubbing: Comment on power et al. *Neuroimage*, 76:436–438.

[23] Chue Hong, N. (2015). Open software for open science.

[24] Churchill, N. W., Oder, A., Abdi, H., Tam, F., Lee, W., Thomas, C., Ween, J. E., Graham, S. J., and Strother, S. C. (2012). Optimizing preprocessing and analysis pipelines for single-subject fmri. i. standard temporal motion and physiological noise correction methods. *Human brain mapping*, 33(3):609–627.

[25] Churchill, N. W., Spring, R., Afshin-Pour, B., Dong, F., and Strother, S. C. (2015). An automated, adaptive framework for optimizing preprocessing pipelines in task-based functional mri. *PloS one*, 10(7):e0131520.

[26] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.

[27] Cohen, M. S. (1997). Parametric analysis of fmri data using linear systems methods. *Neuroimage*, 6(2):93–103.

[28] Collaboration, O. S. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6):657–660.

[29] Cox, R. W. (1996). Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173.

[30] Dafflon, J., Da Costa, P. F., Váša, F., Monti, R. P., Bzdok, D., Hellyer, P. J., Turkheimer, F., Smallwood, J., Jones, E., and Leech, R. (2020). Neuroimaging: into the multiverse. *bioRxiv*.

[31] David, S. P., Ware, J. J., Chu, I. M., Loftus, P. D., Fusar-Poli, P., Radua, J., Munafo, M. R., and Ioannidis, J. P. (2013). Potential reporting bias in fmri studies of the brain. *PloS one*, 8(7):e70104.

[32] Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667.

[33] Dora, L., Agrawal, S., Panda, R., and Abraham, A. (2017). State-of-the-art methods for brain tissue segmentation: A review. *IEEE reviews in biomedical engineering*, 10:235–249.

[34] Dronkers, N. F., Plaisant, O., Iba-Zizen, M. T., and Cabanis, E. A. (2007). Paul broca's historic cases: high resolution mr imaging of the brains of leborgne and lelong. *Brain*, 130(5):1432–1441.

[35] Drummond, C. (2009). Replicability is not reproducibility: nor is it good science.

[36] Duffau, H. (2018). The error of broca: from the traditional localizationist concept to a connectomal anatomy of human brain. *Journal of chemical neuroanatomy*, 89:73–81.

[37] Durnez, J., Degryse, J., Moerkerke, B., Seurinck, R., Sochat, V., Poldrack, R. A., and Nichols, T. E. (2016). Power and sample size calculations for fmri studies based on the prevalence of active peaks. *BioRxiv*, page 049429.

[38] Eaton, J. W., Bateman, D., Hauberg, S., and Wehbring, R. (2017). *GNU Octave version 4.2.1 manual: a high-level interactive language for numerical computations*.

[39] Eglen, S. J., Marwick, B., Halchenko, Y. O., Hanke, M., Sufi, S., Gleeson, P., Silver, R. A., Davison, A. P., Lanyon, L., Abrams, M., et al. (2017). Toward standard practices for sharing computer code and programs in neuroscience. *Nature neuroscience*, 20(6):770–773.

[40] Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28):7900–7905.

[41] Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., and Hariri, A. R. (2020). What is the test-retest reliability of common task-functional mri measures? new empirical evidence and a meta-analysis. *Psychological Science*, 31(7):792–806.

[42] Essig, M., Shiroishi, M. S., Nguyen, T. B., Saake, M., Provenzale, J. M., Enterline, D., Anzalone, N., Dörfler, A., Rovira, À., Wintermark, M., et al. (2013). Perfusion mri: the five most frequently asked technical questions. *AJR. American journal of roentgenology*, 200(1):24.

[43] Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., et al. (2019). fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116.

[44] Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Glasser, M. F., Miller, K. L., Ugurbil, K., and Yacoub, E. (2010). Multiplexed echo planar imaging for sub-second whole brain fmri and fast diffusion imaging. *PloS one*, 5(12):e15710.

[45] Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, 167:104–120.

[46] Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., et al. (2017). Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170.

[47] Fortin, J.-P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., Shinohara, R. T., Initiative, A. D. N., et al. (2016). Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage*, 132:198–212.

[48] Friedman, L., Glover, G. H., Krenz, D., Magnotta, V., and BIRN, T. F. (2006). Reducing inter-scanner variability of activation in a multicenter fmri study: role of smoothness equalization. *Neuroimage*, 32(4):1656–1668.

[49] Friston, K. J., Glaser, D. E., Henson, R. N., Kiebel, S., Phillips, C., and Ashburner, J. (2002a). Classical and bayesian inference in neuroimaging: applications. *Neuroimage*, 16(2):484–512.

[50] Friston, K. J., Jezzard, P., and Turner, R. (1994). Analysis of functional mri time-series. *Human brain mapping*, 1(2):153–171.

[51] Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002b). Classical and bayesian inference in neuroimaging: theory. *NeuroImage*, 16(2):465–483.

[52] Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., and Turner, R. (1996). Movement-related effects in fmri time-series. *Magnetic resonance in medicine*, 35(3):346–355.

[53] Friston, K. J., Zarahn, E., Josephs, O., Henson, R. N., and Dale, A. M. (1999). Stochastic designs in event-related fmri. *Neuroimage*, 10(5):607–619.

[54] Gagnon-Bartsch, J. A., Jacob, L., and Speed, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, pages 1–112.

[55] Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552.

[56] Ganz, M., Nørgaard, M., Beliveau, V., Svarer, C., Knudsen, G. M., and Greve, D. N. (2020). False positive rates in positron emission tomography (pet) voxelwise analyses. *Journal of Cerebral Blood Flow & Metabolism*, page 0271678X20974961.

[57] Gilmore, R. O., Diaz, M. T., Wyble, B. A., and Yarkoni, T. (2017). Progress toward openness, transparency, and reproducibility in cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1396(1):5–18.

[58] Glatard, T., Lewis, L. B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., Rioux, P., Rousseau, M.-E., Sherif, T., Deelman, E., et al. (2015). Reproducibility of neuroimaging analyses across operating systems. *Frontiers in neuroinformatics*, 9:12.

[59] Gonçalves, S. I., De Munck, J. C., Pouwels, P. J., Schoonhoven, R., Kuijer, J. P., Maurits, N. M., Hoogduin, J. M., Van Someren, E. J., Heethaar, R. M., and Da Silva, F. L. (2006). Correlating the alpha rhythm to bold using simultaneous eeg/fmri: inter-subject variability. *Neuroimage*, 30(1):203–213.

[60] Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine*, 8(341):341ps12–341ps12.

[61] Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., and Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, 5:13.

[62] Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data*, 3(1):1–9.

[63] Gorgolewski, K. J. and Poldrack, R. A. (2016). A practical guide for improving transparency and reproducibility in neuroimaging research. *PLoS biology*, 14(7):e1002506.

[64] Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., Sochat, V. V., Nichols, T. E., Poldrack, R. A., Poline, J.-B., et al. (2015). Neurovault. org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in neuroinformatics*, 9:8.

[65] Goutte, C., Nielsen, F. A., and Hansen, K. (2000). Modeling the hemodynamic response in fmri using smooth fir filters. *IEEE transactions on medical imaging*, 19(12):1188–1201.

[66] Gronenschild, E. H., Habets, P., Jacobs, H. I., Mengelers, R., Rozendaal, N., Van Os, J., and Marcelis, M. (2012). The effects of freesurfer version, workstation type, and macintosh operating system version on anatomical volume and cortical thickness measurements. *PloS one*, 7(6):e38234.

[67] Grootoonk, S., Hutton, C., Ashburner, J., Howseman, A., Josephs, O., Rees, G., Friston, K. J., and Turner, R. (2000). Characterization and correction of interpolation effects in the realignment of fmri time series. *NeuroImage*, 11(1):49–57.

[68] Halchenko, Y. O. and Hanke, M. (2012). Open is not enough. let's take the next step: an integrated, community-driven computing platform for neuroscience. *Frontiers in neuroinformatics*, 6:22.

[69] Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. (1993). Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2):413.

[70] Handwerker, D. A., Ollinger, J. M., and D'Esposito, M. (2004). Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, 21(4):1639–1651.

[71] Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W., and Stadler, J. (2014). A high-resolution 7-tesla fmri dataset from complex natural stimulation with an audio movie. *Scientific data*, 1(1):1–18.

[72] Hayasaka, S. and Nichols, T. E. (2003). Validating cluster size inference: random field and permutation methods. *Neuroimage*, 20(4):2343–2356.

[73] Henson, R., Buechel, C., Josephs, O., and Friston, K. (1999). The slice-timing problem in event-related fmri. *NeuroImage*, 9:125.

[74] Herrick, R., Horton, W., Olsen, T., McKay, M., Archie, K. A., and Marcus, D. S. (2016). Xnat central: Open sourcing imaging research data. *NeuroImage*, 124:1093–1096.

[75] Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., and Boulesteix, A.-L. (2021). The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. *Royal Society Open Science*, 8(4):201925.

[76] Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., and Turner, R. (2002). Image distortion correction in fmri: a quantitative evaluation. *Neuroimage*, 16(1):217–240.

[77] Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124.

[78] Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., and David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in cognitive sciences*, 18(5):235–241.

[79] Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., et al. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691.

[80] Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *Neuroimage*, 62(2):782–790.

[81] Jezzard, P. and Balaban, R. S. (1995). Correction for geometric distortion in echo planar images from b0 field variations. *Magnetic resonance in medicine*, 34(1):65–73.

[82] Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.

[83] Johnstone, T., Ores Walsh, K. S., Greischar, L. L., Alexander, A. L., Fox, A. S., Davidson, R. J., and Oakes, T. R. (2006). Motion correction and the use of motion covariates in multiple-subject fmri analysis. *Human brain mapping*, 27(10):779–788.

[84] Kiar, G., Chatelain, Y., de Oliveira Castro, P., Petit, E., Rokem, A., Varoquaux, G., Misic, B., Evans, A. C., and Glatard, T. (2020). Numerical instabilities in analytical pipelines lead to large and meaningful variability in brain networks. *bioRxiv*, pages 2020–10.

[85] Kiar, G., Chatelain, Y., de Oliveira Castro, P., Petit, E., Rokem, A., Varoquaux, G., Misic, B., Evans, A. C., and Glatard, T. (2021). Numerical uncertainty in analytical pipelines lead to impactful variability in brain networks. *PloS one*, 16(11):e0250755.

[86] Kim, Y.-M., Poline, J.-B., and Dumas, G. (2018). Experimenting with reproducibility: a case study of robustness in bioinformatics. *GigaScience*, 7(7):giy077.

[87] Kiss, M., Hermann, P., Vidnyánszky, Z., and Gál, V. (2018). Reducing task-based fmri scanning time using simultaneous multislice echo planar imaging. *Neuroradiology*, 60(3):293–302.

[88] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., et al. (2016). *Jupyter Notebooks-a publishing format for reproducible computational workflows.*, volume 2016.

[89] LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L. K., Yacoub, E., Hu, X., Rottenberg, D., et al. (2003). The evaluation of preprocessing choices in single-subject bold fmri using npairs performance metrics. *NeuroImage*, 18(1):10–27.

[90] Laird, A. R. (2021). Large, open datasets for human connectomics research: Considerations for reproducible and responsible data use. *NeuroImage*, 244:118579.

[91] Lange, N. and Zeger, S. L. (1997). Non-linear fourier time series analysis for human brain mapping by functional magnetic resonance imaging. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(1):1–29.

[92] Lee, M. H., Smyser, C. D., and Shimony, J. S. (2013). Resting-state fmri: a review of methods and clinical applications. *American Journal of neuroradiology*, 34(10):1866–1872.

[93] Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883.

[94] Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161.

[95] Lindquist, M. A., Loh, J. M., Atlas, L. Y., and Wager, T. D. (2009). Modeling the hemodynamic response function in fmri: efficiency, bias and mis-modeling. *Neuroimage*, 45(1):S187–S198.

[96] Liu, T. T. (2012). The development of event-related fmri designs. *Neuroimage*, 62(2):1157–1162.

[97] Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fmri signal. *nature*, 412(6843):150–157.

[98] Lund, T. E., Nørgaard, M. D., Rostrup, E., Rowe, J. B., and Paulson, O. B. (2005). Motion or activity: their role in intra-and inter-subject variation in fmri. *Neuroimage*, 26(3):960–964.

[99] Mackin, D., Fave, X., Zhang, L., Fried, D., Yang, J., Taylor, B., Rodriguez-Rivera, E., Dodge, C., Jones, A. K., and Court, L. (2015). Measuring ct scanner variability of radiomics features. *Investigative radiology*, 50(11):757.

[100] Maclaren, J., Herbst, M., Speck, O., and Zaitsev, M. (2013). Prospective motion correction in brain imaging: a review. *Magnetic resonance in medicine*, 69(3):621–636.

[101] Manoach, D. S., Halpern, E. F., Kramer, T. S., Chang, Y., Goff, D. C., Rauch, S. L., Kennedy, D. N., and Gollub, R. L. (2001). Test-retest reliability of a functional mri working memory paradigm in normal and schizophrenic subjects. *American Journal of Psychiatry*, 158(6):955–958.

[102] Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., Hardcastle, N., Wexler, J., Esteban, O., Goncavles, M., et al. (2021). The openneuro resource for sharing of neuroscience data. *Elife*, 10:e71774.

[103] McNutt, M. (2014). Reproducibility.

[104] Milham, M. P. (2012). Open neuroscience solutions for the connectome-wide association era. *Neuron*, 73(2):214–218.

[105] Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C. E., Morey, R. A., Flashman, L. A., et al. (2016). Inter-site and inter-scanner diffusion mri data harmonization. *NeuroImage*, 135:311–323.

[106] Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., and Uğurbil, K. (2010). Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri. *Magnetic resonance in medicine*, 63(5):1144–1153.

[107] Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2005). The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869.

[108] Mumford, J. A. and Nichols, T. (2006). Modeling and inference of multisubject fmri data. *IEEE Engineering in Medicine and Biology Magazine*, 25(2):42–51.

[109] Mumford, J. A. and Nichols, T. (2009). Simple group fmri modeling and inference. *Neuroimage*, 47(4):1469–1475.

[110] Mumford, J. A., Poline, J.-B., and Poldrack, R. A. (2015). Orthogonalization of regressors in fmri models. *PloS one*, 10(4):e0126255.

[111] Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature human behaviour*, 1(1):1–9.

[112] Nasrallah, I. and Dubroff, J. (2013). An overview of pet neuroimaging. In *Seminars in nuclear medicine*, volume 43, pages 449–461. Elsevier.

[113] Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., Kriegeskorte, N., Milham, M. P., Poldrack, R. A., Poline, J.-B., et al. (2017). Best practices in data analysis and sharing in neuroimaging using mri. *Nature neuroscience*, 20(3):299–303.

[114] Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25.

[115] Nørgaard, M., Ganz, M., Svarer, C., Frokjaer, V. G., Greve, D. N., Strother, S. C., and Knudsen, G. M. (2020). Different preprocessing strategies lead to different conclusions: a [11c] dasb-pet reproducibility study. *Journal of Cerebral Blood Flow & Metabolism*, 40(9):1902–1911.

[116] Oakes, T. R., Johnstone, T., Walsh, K. O., Greischar, L. L., Alexander, A. L., Fox, A. S., and Davidson, R. J. (2005). Comparison of fmri motion correction software tools. *Neuroimage*, 28(3):529–543.

[117] Pauli, R., Bowring, A., Reynolds, R., Chen, G., Nichols, T. E., and Maumet, C. (2016). Exploring fmri results space: 31 variants of an fmri analysis in afni, fsl, and spm. *Frontiers in neuroinformatics*, 10:24.

[118] Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., and Nichols, T. E. (2011). *Statistical parametric mapping: the analysis of functional brain images*. Elsevier.

[119] Pernet, C. and Poline, J.-B. (2015). Improving functional magnetic resonance imaging reproducibility. *Gigascience*, 4(1):1–8.

[120] Petersen, S. E. and Dubis, J. W. (2012). The mixed block/event-related design. *Neuroimage*, 62(2):1177–1184.

[121] Phillips, A. A., Chan, F. H., Zheng, M. M. Z., Krassioukov, A. V., and Ainslie, P. N. (2016). Neurovascular coupling in humans: physiology, methodological advances and clinical implications. *Journal of Cerebral Blood Flow & Metabolism*, 36(4):647–664.

[122] Poldrack, R. A. (2013). Anatomy of a coding error.

[123] Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., and Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature reviews neuroscience*, 18(2):115–126.

[124] Poldrack, R. A., Barch, D. M., Mitchell, J., Wager, T., Wagner, A. D., Devlin, J. T., Cumba, C., Koyejo, O., and Milham, M. (2013). Toward open sharing of task-based fmri data: the openfmri project. *Frontiers in neuroinformatics*, 7:12.

[125] Poldrack, R. A. and Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nature neuroscience*, 17(11):1510–1517.

[126] Poldrack, R. A., Laumann, T. O., Koyejo, O., Gregory, B., Hover, A., Chen, M.-Y., Gorgolewski, K. J., Luci, J., Joo, S. J., Boyd, R. L., et al. (2015). Long-term neural and physiological phenotyping of a single human. *Nature communications*, 6(1):1–15.

[127] Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.

[128] Poline, J.-B., Breeze, J. L., Ghosh, S. S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., Helmer, K. G., Marcus, D. S., Poldrack, R. A., Schwartz, Y., et al. (2012). Data sharing in neuroimaging research. *Frontiers in neuroinformatics*, 6:9.

[129] Posner, M. I., Petersen, S. E., Fox, P. T., and Raichle, M. E. (1988). Localization of cognitive operations in the human brain. *Science*, 240(4859):1627–1631.

[130] Rademacher, J., Caviness Jr, V., Steinmetz, H., and Galaburda, A. (1993). Topographical variation of the human primary cortices: implications for neuroimaging, brain mapping, and neurobiology. *Cerebral Cortex*, 3(4):313–329.

[131] Rolland, X., Maurel, P., and Maumet, C. (2022). Towards efficient fmri data re-use: Can we run between-group analyses with datasets processed differently with spm? In *IEEE International Symposium on Biomedical Imaging (ISBI 2022))*.

[132] Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638.

[133] Roy, C. S. and Sherrington, C. S. (1890). On the regulation of the blood-supply of the brain. *The Journal of physiology*, 11(1-2):85–158.

[134] Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D., and Nichols, T. E. (2009). Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *Neuroimage*, 45(3):810–823.

[135] Schadt, E. E., Li, C., Ellis, B., and Wong, W. H. (2001). Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, 84(S37):120–125.

[136] Schloss, P. D. (2018). Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *MBio*, 9(3):e00525–18.

[137] Setsompop, K., Gagoski, B. A., Polimeni, J. R., Witzel, T., Wedeen, V. J., and Wald, L. L. (2012). Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magnetic resonance in medicine*, 67(5):1210–1224.

[138] Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press.

[139] Shaw, M. E., Strother, S. C., Gavrilescu, M., Podzebenko, K., Waites, A., Watson, J., Anderson, J., Jackson, G., and Egan, G. (2003). Evaluating subject specific preprocessing choices in multisubject fmri data sets using data-driven performance metrics. *NeuroImage*, 19(3):988–1001.

[140] Sladky, R., Friston, K. J., Tröstl, J., Cunnington, R., Moser, E., and Windischberger, C. (2011). Slice-timing effects and their correction in functional mri. *Neuroimage*, 58(2):588–594.

[141] Smith, A. M., Lewis, B. K., Ruttimann, U. E., Frank, Q. Y., Sinnwell, T. M., Yang, Y., Duyn, J. H., and Frank, J. A. (1999). Investigation of low frequency drift in fmri signal. *Neuroimage*, 9(5):526–533.

[142] Smith, S. M. (2004). Overview of fmri analysis. *The British Journal of Radiology*, 77(suppl_2):S167–S175.

[143] Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the national academy of sciences*, 106(31):13040–13045.

[144] Stehling, M. K., Turner, R., and Mansfield, P. (1991). Echo-planar imaging: magnetic resonance imaging in a fraction of a second. *Science*, 254(5028):43–50.

[145] Strother, S., La Conte, S., Hansen, L. K., Anderson, J., Zhang, J., Pulapura, S., and Rottenberg, D. (2004). Optimizing the fmri data-processing pipeline using prediction and reproducibility performance metrics: I. a preliminary group analysis. *Neuroimage*, 23:S196–S207.

[146] Strother, S. C. (2006). Evaluating fmri preprocessing pipelines. *IEEE Engineering in Medicine and Biology Magazine*, 25(2):27–41.

[147] Szucs, D. and Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS biology*, 15(3):e2000797.

[148] Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., and Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS one*, 6(6):e21101.

[149] Thompson, P. M., Schwartz, C., Lin, R. T., Khan, A. A., and Toga, A. W. (1996). Three-dimensional statistical analysis of sulcal variability in the human brain. *Journal of Neuroscience*, 16(13):4261–4274.

[150] Tseng, G. C., Oh, M.-K., Rohlin, L., Liao, J. C., and Wong, W. H. (2001). Issues in cdna microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic acids research*, 29(12):2549–2557.

[151] Uri, Joe, and Leif (2021). Evidence of fraud in an influential field experiment about dishonesty.

[152] Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79.

[153] Van Horn, J. D. and Gazzaniga, M. S. (2013). Why share data? lessons learned from the fmridc. *Neuroimage*, 82:677–682.

[154] Van Horn, J. D., Grafton, S. T., Rockmore, D., and Gazzaniga, M. S. (2004). Sharing neuroimaging studies of human cognition. *Nature neuroscience*, 7(5):473–481.

[155] Van Horn, J. D., Grethe, J. S., Kostelec, P., Woodward, J. B., Aslam, J. A., Rus, D., Rockmore, D., and Gazzaniga, M. S. (2001). The functional magnetic resonance imaging data center (fmridc): the challenges and rewards of large–scale databasing of neuroimaging studies. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1323–1339.

[156] Varoquaux, G. (2018). Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage*, 180:68–77.

[157] Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. *Perspectives on psychological science*, 4(3):274–290.

[158] Whitaker, K. (2016). Making your research reproducible.

[159] Wicherts, J. M., Bakker, M., and Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS one*, 6(11):e26828.

[160] Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2004a). Multilevel linear modelling for fmri group analysis using bayesian inference. *Neuroimage*, 21(4):1732–1747.

[161] Woolrich, M. W., Behrens, T. E., and Smith, S. M. (2004b). Constrained linear basis sets for hrf modelling using variational bayes. *NeuroImage*, 21(4):1748–1761.

[162] Woolrich, M. W., Ripley, B. D., Brady, M., and Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fmri data. *Neuroimage*, 14(6):1370–1386.

[163] Xu, J., Moeller, S., Strupp, J., Auerbach, E., Chen, L., Feinberg, D., Ugurbil, K., and Yacoub, E. (2012). Highly accelerated whole brain imaging using aligned-blipped-controlled-aliasing multiband epi. In *Proceedings of the 20th Annual Meeting of ISMRM*, volume 2306.

Publications

Open research: linking the bits and pieces with OpenAIRE-connect
Camille Maumet, Xavier Rolland, Axel Bonnet, Sorina Camarasu-Pop, Argiro Kokogiannaki, Christian Barillot
OHBM 2019-25th Annual Meeting of the Organization for Human Brain Mapping. 2019. p. 1-6.

Towards Efficient FMRI Data Re-Use: Can We Run Between-Group Analyses with Datasets Processed Differently with SPM?
Xavier Rolland, Pierre Maurel, Camille Maumet
2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). IEEE, 2022. p. 1-4.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre : Impact de la Variabilité Analytique sur la Compatibilité Entre les Données dans le Etudes d'Imagerie par Résonance Magnétique Fonctionnelle

Mot clés : Variabilité Analytique, Reproductibilité, IRMf de tâche, Hypothèse Nulle, Compatibilité Entre Les Données

Résumé : Au cours des dernières années, le manque de reproductibilité des résultats de recherche est devenu un sujet majeur dans de nombreux domaines scientifiques, y compris l'Imagerie par Résonance Magnétique Fonctionnelle (IRMf). La faible puissance statistique liée aux petites tailles d'échantillons a été identifiée comme l'un des principaux facteurs de non-reproductibilité dans les études d'IRMf. Le développement du partage de données en neuroimagerie ouvre de nouvelles opportunités, permettant d'effectuer des études avec des tailles d'échantillons plus grandes en réutilisant des données existantes, provenant éventuellement d'ensembles de données différents. Cependant, cela peut conduire à combiner des données traitées différemment.

Dans cette thèse, nous avons étudié l'impact de la variabilité causée par les différences entre les chaînes de traitement, appelée variabilité analytique, sur la validité de nouvelles analyses lorsque l'on combine des données de sujet traitées avec différentes chaînes de traitement au niveau individuel. Les analyses ont été effectuées avec des chaînes de traitement qui différaient les unes des autres sur un ensemble de paramètres définis. Nous avons observé que la variabilité induite par les différences de valeurs pour ces paramètres entre les chaînes de traitement était acceptable dans certains cas et rédhibitoire dans d'autres cas. Nous avons conclu que les différences en terme de traitement appliqué sur les données de sujet doivent être prises en compte avant de combiner ces données. Enfin, nous avons proposé une méthode de correction de l'effet de la variabilité analytique pour ces analyses de groupe.

Title: Impact of Analytical Variability on Data Compatibility in Functional Magnetic Resonance Imaging Studies

Keywords: Analytical Variability, Reproducibility, fMRI, Null Hypothesis, Data Compatibility

Abstract: In recent years, the lack of reproducibility of research findings has become an important source of concern in many scientific fields, including functional Magnetic Resonance Imaging (fMRI). The low statistical power induced by low sample sizes was identified as one of the leading causes of irreproducibility in fMRI studies. The development of data sharing in the field of neuroimaging opens up new opportunities to perform studies with larger sample sizes by reusing existing data, possibly coming from different datasets. However, doing so may require combining data which have been processed differently.

In this thesis, we investigated the impact of analytical variability – the variability induced by different processing pipelines – on the validity of new analyses when combining subject data processed with different subject-level pipelines. Analyses were performed using pipelines that differed from each other on a set of given parameters. We found that the analytical variability induced by the parameter differences between pipelines was acceptable for some of these analyses and redhibitory for others. We concluded that differences in processing applied on subject data have to be taken into account before combining these data. Finally, a first steps toward the correction of the effect of analytical variability for these between-group analyses is proposed.