Université de Lille

École Doctorale Biologie Santé de Lille

Thèse de doctorat

Spécialité *Recherche clinique, innovation technologique, santé publique*

Harmonisation multicentrique d'images IRM du cerveau avec des modèles génératifs non-supervisés

par Vincent Roca

Soutenue publiquement le 19 décembre 2023

Composition du jury

Michel Dojat — *rapporteur*
directeur de recherche INSERM, Université Grenoble Alples

Nicolas Menjot de Champfleur — *rapporteur*
professeur des universités, praticien hospitalier, CHU de Montpellier

Christine Fernandez-Maloigne — *examinatrice*
professeur des universités, Université de Poitiers

Jean-Pierre Pruvo, président du Jury — *examinateur*
professeur des universités, praticien hospitalier, CHU de Lille

Renaud Lopes — *directeur de thèse*
maître de conférences des universités, praticien hospitalier, CHU de Lille

Grégory Kuchcinski — *co-encadrant de thèse*
maître de conférences des universités, praticien hospitalier, CHU de Lille

Résumé

L'imagerie par résonance magnétique (IRM) permet l'acquisition d'images du cerveau pour l'étude de maladies neurologiques et psychiatriques. Les images IRM sont de plus en plus utilisées dans des études statistiques pour identifier des biomarqueurs et pour des modèles de prédiction. Pour gagner en puissance statistique, ces études agrègent parfois des données acquises avec différentes machines, ce qui peut introduire de la variabilité technique biaisant les analyses des variabilités biologiques.

Ces dernières années, des méthodes d'harmonisation ont été proposées pour limiter l'impact de ces variabilités dans les analyses. De nombreuses études ont notamment travaillé sur des modèles génératifs basés sur de l'apprentissage profond non-supervisé. Le travail de thèse s'inscrit dans le cadre de ces modèles qui constituent un champ de recherche prometteur mais encore exploratoire.

Dans la première partie de ce manuscrit, une revue des méthodes d'harmonisation rétrospective est proposée. Différentes méthodes de normalisation appliquées au niveau de l'image, de translation de domaines ou de transfert de style y sont décrites en vue de comprendre leurs enjeux respectifs, avec une attention particulière portée aux modèles génératifs non-supervisés.

La deuxième partie porte sur les méthodes d'évaluation de l'harmonisation rétrospective. Une revue de ces méthodes est d'abord réalisée. Les plus communes reposent sur des sujets "voyageurs" pour présumer des vérités terrain à l'harmonisation. La revue présente également des évaluations employées en l'absence de tels sujets : étude de différences inter-domaine, de motifs biologiques et de performances de modèles prédictifs. Des expériences mettant en avant des limites de certaines approches couramment employées et des points d'attention nécessaires à leur utilisation sont ensuite proposées.

La troisième partie présente un nouveau modèle d'harmonisation d'images IRM cérébrales basé sur une architecture CycleGAN. Contrairement aux précédents travaux, le modèle est tridimensionnel et traite les volumes complets. Des images IRM provenant de six jeux de données variables en termes de paramètres d'acquisition et de distribution d'âge sont utilisées pour expérimenter la méthode. Des analyses de distributions d'intensités, de volumes cérébraux, de métriques de qualité d'image et de caractéristiques radiomiques montrent une homogénéisation efficace entre les différents sites de l'étude. À côté de ça, la conservation et le renforcement de motifs biologiques sont montrés avec une analyse de l'évolution d'estimations de volumes de matière grise avec l'âge, des expériences de prédiction d'âge, la cotation de motifs radiologiques dans les images et une évaluation supervisée avec un jeu de données de sujets voyageurs.

La quatrième partie présente également une méthode d'harmonisation originale avec des modifications majeures de la première en vue d'établir un générateur "universel" capable d'harmoniser des images sans connaître leur domaine d'origine. Après un entraînement exploitant des données acquises avec onze scanners IRM, des expériences sur des images de sites non-vus lors de l'entraînement montrent un renforcement de motifs cérébraux liés à l'âge et à la maladie d'Alzheimer après harmonisation. De plus, des comparaisons avec d'autres approches d'harmonisation d'intensités suggèrent que le modèle est plus efficace et plus robuste dans différentes tâches subséquentes à l'harmonisation.

Ces différents travaux constituent une contribution significative au domaine de l'harmonisation rétrospective d'images IRM cérébrales. Les documentations bibliographiques fournissent en effet un corpus de connaissances méthodologiques pour les

futurs études dans ce domaine, que ce soit pour l'harmonisation en elle-même ou pour la validation. De plus, les deux modèles développés sont deux outils robustes accessibles publiquement qui pourraient être intégrés à de futures études multicentriques en IRM.

Abstract

Magnetic resonance imaging (MRI) enables the acquisition of brain images used in the study of neurologic and psychiatric diseases. MR images are more and more used in statistical studies to identify biomarkers and for predictive models. To improve statistical power, these studies sometimes pool data acquired with different machines, which may introduce technical variability and bias into the analysis of biological variabilities.

In the last few years, harmonization methods have been proposed to limit the impact of these variabilities. Many studies have notably worked on generative models based on unsupervised deep learning. The doctoral research is within the context of these models, which constitute a promising but still exploratory research field.

In the first part of this manuscript, a review of the prospective harmonization methods is proposed. Different methods consisting in normalization applied at the image level, domain translation or style transfer are described to understand their respective issues, with a special focus on unsupervised generative models.

The second part is about the methods for evaluation of retrospective harmonization. A review of these methods is first conducted. The most common rely on “traveling” subjects to assume ground truths for harmonization. The review also presents evaluations employed in the absence of such subjects: study of inter-domain differences, biological patterns and performances of predictive models. Experiments showing limits of some approaches commonly employed and important points to consider for their use are then proposed.

The third part presents a new model for harmonization of brain MR images based on a CycleGAN architecture. In contrast with the previous works, the model is three-dimensional and processes full volumes. MR images from six datasets that vary in terms of acquisition parameters and age distributions are used to test the method. Analyses of intensity distributions, brain volumes, image quality metrics and radiomic features show an efficient homogenisation between the different sites of the study. Next, the conservation and the reinforcement of biological patterns are demonstrated with an analysis of the evolution of gray-matter volume estimations with age, experiments of age prediction, ratings of radiologic patterns in the images and a supervised evaluation with a traveling subject dataset.

The fourth part also presents an original harmonization method with major updates of the first one in order to establish a “universal” generator able to harmonize images without knowing their domain of origin. After a training with data acquired on eleven MRI scanners, experiments on images from sites not seen during the training show a reinforcement of brain patterns relative to age and Alzheimer after harmonization. Moreover, comparisons with other intensity harmonization approaches suggest that the model is more efficient and more robust to different tasks subsequent to harmonization.

These different works are a significant contribution to the domain of retrospective harmonization of brain MR images. The bibliographic documentations indeed provide a methodological knowledge base for the future studies in this domain, whether for harmonization in itself or for validation. In addition, the two developed models are two robust tools publicly available that may be integrated in future MRI multicenter studies.

Remerciements

À M. Michel Dojat et M. Nicolas Menjot de Champfleur, je vous remercie d'avoir participé à mes comités de suivi de thèse. Les échanges que nous y avons eus m'ont été profitables pour l'avancée de mon travail. Je vous suis également reconnaissant d'avoir accepté d'être rapporteur de cette thèse et j'espère que vous trouverez de l'intérêt dans sa lecture.

À Mme Cristine Fenandez-Maloigne, je vous remercie d'avoir accepté de participer à cette soutenance pour examiner mon travail de thèse.

À M. Jean-Pierre Pruvo, je vous remercie d'avoir accepté d'être examinateur et Président du Jury pour cette soutenance. Je suis reconnaissant du travail que vous avez accompli pour développer la recherche en IRM au CHU de Lille et la plateforme du Liife.

À Renaud Lopes, je te remercie de m'avoir fait confiance pour ce projet de thèse. Tes remarques sur mon travail, tes conseils et ton expérience dans la recherche ont été indispensables pour la réussite de ce projet.

À Grégory Kuchcinski, je te remercie également pour les retours critiques que tu as faits sur mon travail, ainsi que pour l'expertise radiologique et scientifique que tu as apportée dans ce projet.

Je remercie aussi toutes les personnes grâce auxquelles j'ai pris plaisir à évoluer au sein du laboratoire du Liife durant ces trois années : Romain, Cécile, Julien, Maxime, Jean-Baptiste, Sabine, Amal, Morgan, Dorian, Félix.

Enfin, je suis reconnaissant envers les Plateformes Lilloises en Biologie Santé (PLBS) et le Lille Invivo Imaging and Functional Exploration (LIIFE) qui sont les structures auxquelles j'étais rattaché pour ce doctorat. Je suis également reconnaissant envers le CHU de Lille et la société Philips, dont le partenariat a donné lieu au financement de cette thèse.

Table des matières

Liste des abréviations

ACG : atrophie corticale globale
ACP : analyse en composantes principales
AF : anisotropie fractionnelle
ATM : atrophie temporale médiale
CALAMITI : méthode d'harmonisation de Zuo et al. (2021b, 2021a)
CNN : réseau de neurones convolutionnel
DAPM : différence d'âge prédit moyenne
DM : diffusivité moyenne
DME : déviation par rapport à la moyenne d'entraînement
EAM : erreur absolue moyenne
EPD : espace périvasculaire dilaté
EPD-CS : le nombre d'espaces périvasculaires dilatés dans le centre semi-ovale
EPD-GB : le nombre d'espaces périvasculaires dilatés dans le ganglion basal
GAN : réseaux antagonistes génératifs
HM : méthode d'*histogram matching* (Nyul et al. 2000; Shah et al. 2011)
IQM : métrique de qualité d'image
IRM : imagerie par résonance magnétique
KS : Kolmogorov-Smirnov
LCS : liquide cérébrospinal
MB : matière blanche
MG : matière grise
RME : régression vers la moyenne d'entraînement
SSIM : structural similarity index measure
STGAN : méthode d'harmonisation de Liu et al. (2023)
SVM : machine à vecteur de support
T : Tesla
T1w : T1-pondérée
T2w : T2-pondérée
WS : WhiteStripe

1. Introduction

1.1. Imagerie par résonance magnétique

L'imagerie par résonance magnétique (IRM) est une technique médicale qui permet d'obtenir des images de haute résolution des tissus mous humains. L'IRM cérébrale offre une visualisation détaillée de certaines structures, ce qui en fait une approche courante pour étudier certaines pathologies comme la maladie d'Alzheimer, la maladie de Parkinson ou la sclérose en plaques.

Cette technique d'imagerie repose sur un champ magnétique plus ou moins puissant, sur l'émission d'ondes électromagnétiques qui font faire résonner les atomes d'hydrogène et ces derniers émettront des signaux captés par des antennes. Différents types d'IRM existent, par exemple :

- L'IRM structurelle : Elle permet d'obtenir une image en trois dimensions pour la visualisation des structures cérébrales et la détection de lésions.
- L'IRM de diffusion : Elle représente la diffusion de molécules d'eau dans le cerveau. L'imagerie du tenseur de diffusion permet par exemple de visualiser les faisceaux de matière blanche.
- L'IRM fonctionnelle : Elle permet d'enregistrer une activité neuronale indirecte par un couplage neuro-vasculaire, que ce soit au repos ou durant l'accomplissement de certaines tâches spécifiques.

Parmi les IRMs structurelles, les séquences 3D de pondération T1 (T1w) sont les plus utilisées dans les protocoles de recherche en neuroimagerie. La majorité des logiciels d'extraction de caractéristiques reposent sur cette séquence. Elle permet par exemple de segmenter automatiquement la matière grise (MG), la matière blanche (MB) et le liquide cérébrospinal (LCS) (Figure 1), ce qui est utilisé dans diverses études cliniques (Erickson et al. 2014; Gautam et al. 2014; Grieve et al. 2013).

1.2. La variabilité inter-sites en IRM

La richesse des informations qu'elle fournit a fait de l'IRM une technologie très utilisée, notamment en recherche clinique. Toutefois, c'est une technique coûteuse en temps et en argent. Il est donc difficile d'acquérir un grand nombre d'images sur une même machine. En vue d'augmenter la taille des échantillons pour couvrir plus largement les différentes caractéristiques d'une population, beaucoup de projets ont ainsi récemment regroupé des images provenant de multiples sites d'acquisition. On peut par exemple citer Alzheimer's Disease Neuroimaging Initiative (ADNI), Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) ou SRPBS (Tanaka et al. 2021).

Il est cependant connu que l'utilisation de données multicentriques peut introduire des variabilités non biologiques dans les données qui peuvent être liées aux constructeurs, à l'intensité du champ ou à la conception des séquences (Chen et al. 2014; Hawco et al. 2018; Jovicich et al. 2006; Reig et al. 2009; Takao et al. 2011). La Figure 2 illustre avec un exemple des variabilités obtenues avec différentes machines d'acquisition. Cependant, même si des différences de qualité d'image peuvent amener des inégalités dans la précision du diagnostic, ces variabilités ne sont pas vraiment problématiques d'un point de vue radiologique. En revanche, le nombre croissant d'études impliquant des traitements automatiques de données IRM donne de l'importance à ce phénomène. L'hétérogénéité due

à l'acquisition peut en effet mitiger la puissance statistique gagnée en agrégeant des données de multiples sources, voire plus grave, mener à des fausses conclusions scientifiques en cas de confusion entre variabilité biologique et variabilité inter-sites. Malgré les efforts pour la standardisation des acquisitions dans certains projets multicentriques, des différences peuvent toujours être constatées entre les sites (Kruggel et al. 2010; Shinohara et al. 2017).

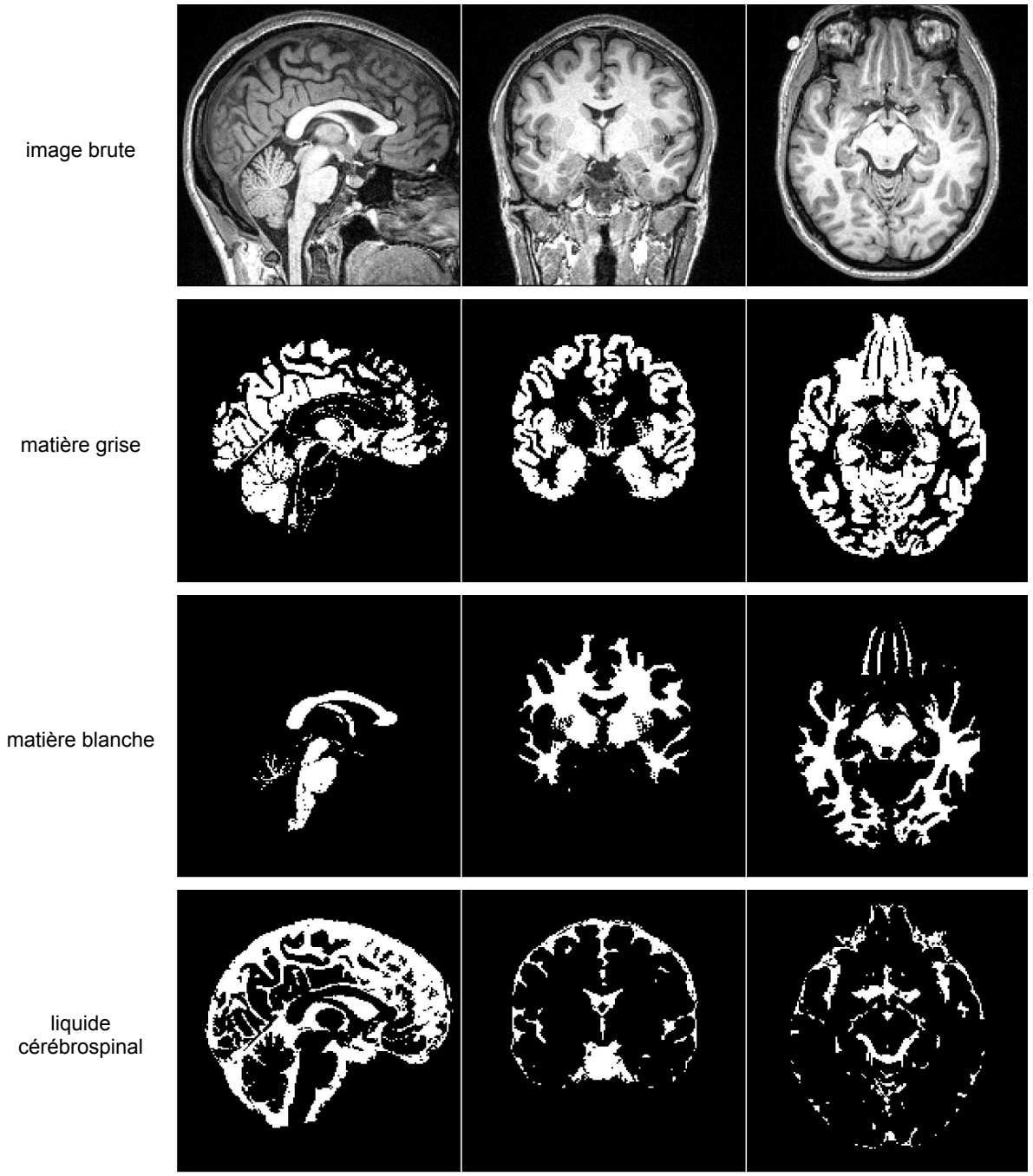


Figure 1 : Exemple d'une segmentation automatisée faite avec SPM12.

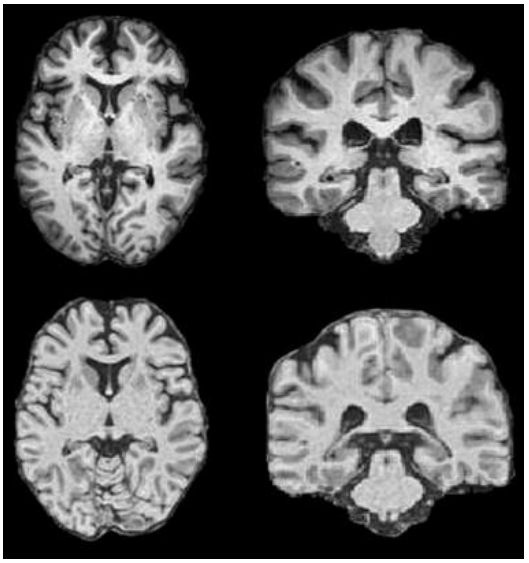


Figure 2 : Images IRM T1-pondérées d'un même sujet acquises avec 2 différentes machines, une Philips Achieva 3T (première ligne) et une General Electric Excite PA 1.5T (deuxième ligne). Figure adaptée de Kruggel et al. (2010).

1.3. L'intelligence artificielle en IRM

On peut définir l'intelligence artificielle (IA) comme l'ensemble des méthodes informatiques visant à résoudre des problèmes avec des solutions non exactes. Les principales méthodes d'IA sont les méthodes d'apprentissage automatique. Elles apprennent la résolution d'un problème par l'exploitation de données de manière automatique. Parmi elles, on distingue d'abord les méthodes d'apprentissage supervisé qui recherchent d'éventuels motifs dans les données associant une *entrée* à un ou plusieurs *labels*. Différentes classifications supervisées ont été mises en place dans des études d'IRM pour la prédiction de diagnostics cliniques en s'appuyant sur l'expertise humaine (Rathore et al. 2017; Sabuncu et Konukoglu 2015; Sinha et al. 2021). La prédiction d'âge a également eu un impact important sur la recherche ces dernières années (Sajedi et Pardakhti 2019). L'idée est d'apprendre à prédire l'âge d'un sujet à partir de données IRM. Des études ont montré que l'âge prédit (communément appelé âge cérébral) pouvait être un biomarqueur associé à diverses pathologies et dysfonctionnements (Bashyam et al. 2020; Cole et al. 2018). On peut également évoquer la problématique plus complexe de la segmentation des IRMs cérébrales qui est un champ de recherche majeur depuis depuis des décennies et pour lequel de nombreux outils ont été mis en place (Despotović et al. 2015). Récemment, des chercheurs ont développé des méthodes d'apprentissage supervisé pour accomplir cette tâche, par exemple pour la localisation de tumeurs (Naser et Deen 2020) ou pour la segmentation du cerveau complet (Huo et al. 2019).

Les modèles d'apprentissage supervisé présentent l'avantage d'être simples et intelligibles pour les humains. En effet, bien que l'entraînement puisse être relativement autonome, il s'agit néanmoins toujours de reproduire l'expertise humaine. Cela peut toutefois présenter quelques inconvénients comme la nécessité de labelliser les données de la base d'entraînement (e.g. diagnostic clinique, segmentation) et les potentiels biais induits par le(s) labellisateur(s). Les approches non supervisées visent à s'affranchir de ces limites tout en continuant d'exploiter les grands volumes de données à disposition. Alors que les méthodes

supervisées prédisent des labels, les méthodes non supervisées identifient des motifs propres aux données et sont plus exploratoires. Ce type d'approche a beaucoup été utilisé ces dernières années sur des données IRM, notamment pour la réduction de dimension (Tang et al. 2021), le clustering (Tan et al. 2022) et la synthèse d'images (Dai et al. 2020).

Parmi les méthodes d'apprentissage automatique, il est également important de distinguer celle qui utilise de l'apprentissage *profond*. L'idée générale de l'apprentissage profond est l'implémentation de multiples transformations sur les données qui permettent d'apprendre des motifs d'un niveau élevé d'abstraction. L'objectif est que les algorithmes soient capables de traiter des données de grande dimension en extrayant *par eux-mêmes* les informations intéressantes pour une tâche donnée. Les modèles implémentés sont appelés réseaux de neurones. L'apprentissage profond a principalement prospéré au cours des dix dernières années avec le traitement d'images et les réseaux de neurones convolutionnels (CNN) (Chauhan et al. 2018; He et al. 2016). Les images IRM diffèrent des images classiques au format RVB car elles n'ont qu'un seul canal d'intensité et sont souvent 3D. La généricité de l'apprentissage profond a toutefois permis de nombreuses applications de CNNs en IRM (Hardaha et al. 2023).

1.4. L'apprentissage automatique pour l'harmonisation d'IRMs

La problématique de la variabilité inter-sites a fait émerger un champ de recherche ces dernières années portant sur l'harmonisation inter-sites, c'est-à-dire les approches qui cherchent à limiter les variabilités qui viennent du processus d'acquisition des IRMs (ou à limiter leurs effets dans les analyses) tout en conservant l'information pertinente (i.e. l'information biologique). Pour aller dans ce sens, une pratique courante dans les protocoles de recherche en IRM consiste à harmoniser les séquences avant la phase d'acquisition des images sur la population d'intérêt. Cela passe souvent par l'utilisation de fantômes ou de quelques sujets volontaires sains qui vont permettre de standardiser au préalable certaines caractéristiques d'imagerie. Dans le cadre des études multicentriques, certaines recommandations sont faites pour l'harmonisation des protocoles (De Stefano et al. 2022) et des chercheurs ont également travaillé sur l'implémentation de séquences et de systèmes de calibration (Clarke et al. 2020).

Ces standardisations des protocoles d'acquisition sont complémentées par des recherches autour de méthodes qui utilisent de l'apprentissage automatique pour harmoniser des données déjà acquises, on parle alors d'harmonisation rétrospective. L'harmonisation rétrospective est nécessaire dans les études multicentriques de par l'impossibilité de garantir l'absence de variabilité inter-sites, même en cas d'efforts vers des méthodes d'acquisition homogènes. La diversité des constructeurs et des caractéristiques techniques des machines est en effet difficilement évitable si l'on souhaite utiliser un jeu de données avec suffisamment de sujets. Par ailleurs, l'émergence de multiples approches rétrospectives a été favorisée par la disponibilité de données IRM issues de multiples sources (par exemple les bases citées en section 1.2) qui couvrent de nombreuses caractéristiques d'acquisition différentes. L'augmentation des puissances de calcul informatique a également aidé au développement de nouveaux modèles capables d'exploiter ces quantités importantes de données. Comme nous verrons dans la section 2, l'utilisation de l'apprentissage automatique pour l'harmonisation inter-sites d'IRMs s'est faite autour d'une variété importante d'approches incluant différents niveaux de supervision et de profondeur des algorithmes.

1.5. Stratégie adoptée

Cette thèse de doctorat s'inscrit dans le contexte des méthodes d'harmonisation rétrospectives pour les études multicentriques. Plus particulièrement, nous nous sommes intéressés à l'utilisation de l'apprentissage profond pour l'harmonisation inter-sites. Beaucoup d'études ont été faites avec ce type d'approche ces dernières années, mais elles constituent un champ de recherche encore très exploratoire et riche en problématiques techniques relatives à l'analyse IRM et plus généralement à l'apprentissage automatique pour l'harmonisation (sections 2 et 3). Le travail de thèse a principalement porté sur le développement de deux nouvelles méthodes pour l'harmonisation d'images IRM structurelles et sur la validation de ces méthodes par différentes expériences incluant des données de multiples sources d'acquisition.

Comme nous le verrons tout au long de ce manuscrit, l'une des problématiques majeures dans le domaine de l'harmonisation inter-sites est le compromis entre l'homogénéisation des données et la conservation de l'information biologique. Cette problématique affecte deux parties du travail de recherche : la première concerne le développement méthodologique et les processus d'optimisation qui doivent favoriser un *bon* équilibre entre les deux aspects ; la deuxième, tout aussi importante, porte sur l'évaluation des méthodes et des données harmonisées qui doit rendre compte à la fois des variabilités inter-sites que l'on cherche à réduire mais également des variabilités inter-individus que l'on cherche à conserver. La thèse s'est donc située autour de ces deux aspects du travail de recherche et a permis d'aborder des problématiques techniques propres à l'apprentissage profond mais également des questions spécifiques à l'IRM cérébrale.

En vue de restituer le travail de recherche, le présent manuscrit suit le plan suivant :

1. **Revue des méthodes d'harmonisation.** Ce chapitre revient en détail sur les différentes approches méthodologiques qui sont utilisées pour limiter la variabilité technique en IRM suite à l'acquisition des images.
2. **Revue des méthodes d'évaluation de l'harmonisation.** Une première partie présente les méthodes qui ont été proposées dans la littérature afin de rendre compte de la qualité des méthodes d'harmonisation. Une deuxième partie présente quelques expériences que nous avons menées pour illustrer certaines problématiques et certaines limites d'évaluations couramment mises en place.
3. **Proposition d'un modèle génératif pour une harmonisation par paire de domaines.** Basé sur un apprentissage profond non-supervisé, le modèle implémenté en 3D est appliqué sur des images cérébrales T1w venant de 6 sites d'acquisition différents. Diverses approches de validation illustrent l'intérêt du modèle pour des études multicentriques.
4. **Proposition d'un générateur universel pour une harmonisation multi-domaine.** Dans ce chapitre, une architecture de modèle originale est mise en place afin de construire un générateur pouvant harmoniser des images IRM venant de sites inconnus. L'apprentissage unifie des translations entre un nombre arbitraire de domaines. Les expériences menées mettent en évidence la préservation et le renforcement de l'information anatomique dans les images suite à l'harmonisation.
5. **Discussion générale et perspectives autour des travaux de thèse.**

2. Revue des méthodes d'harmonisation rétrospective en IRM

2.1. Introduction

Une grande variété de méthodes permet de limiter la variabilité technique en IRM suite à l'acquisition des images. Ces méthodes vont des prétraitements IRM classiques appliqués au niveau individuel à des modèles complexes d'apprentissage profond exploitant de grandes bases de données. Dans ce chapitre, nous en dressons un inventaire exhaustif avec une focalisation sur les aspects méthodologiques.

2.2. Prétraitements du signal IRM

Une première catégorie d'approches regroupe des outils de pré-traitement automatiques communément utilisés dans divers domaines de recherche sur l'IRM. Ces outils sont particulièrement utiles lorsque l'on traite des jeux de données multicentriques (Li et al. 2021). Ils servent même souvent d'étape préalable aux méthodes d'harmonisation plus poussées et ont un double objectif : éliminer des variabilités techniques simples et faciliter les futurs traitements. Trois prétraitements typiques sont la correction des inhomogénéités de champ, le recalage dans un espace commun et l'extraction de la boîte crânienne.

2.2.1. Correction d'inhomogénéités de champ

La correction d'inhomogénéités vise à éliminer des variabilités de basse fréquence dans l'image liées au processus d'acquisition. L'algorithme N3 (Sled et al. 1998) a pendant longtemps été l'outil standard pour la correction automatique d'inhomogénéités. Certains chercheurs l'ont utilisé récemment comme prétraitement à leur outil d'harmonisation (Chen et al. 2021; Liu et al. 2023). Cependant, N3 a largement été remplacé depuis une dizaine d'années par l'algorithme N4ITK (Tustison et al. 2010), plus robuste, plus rapide et plus souvent intégré dans les pipelines d'harmonisation (Dewey et al. 2019, 2020; Pomponio et al. 2020; Zuo et al. 2021b). Un exemple d'application de N4ITK est montré dans la Figure 3.

2.2.2. Recalage dans un espace commun

Le recalage des images IRM dans un espace commun consiste en un alignement des régions anatomiques dans des localisations similaires de la matrice de voxels, de telle sorte qu'un même voxel correspond à la même région dans différentes images. Le système de coordonnées commun permet par exemple d'analyser des variabilités d'intensité par région à travers différentes images. Outre l'alignement, le recalage permet de standardiser les dimensions et la résolution des images, ce qui peut être particulièrement important pour les analyses basées sur des réseaux de neurones.

Pour le prétraitement à l'harmonisation, la pratique commune est le recalage linéaire conservant la structure du cerveau (Bashyam et al. 2022; Cackowski et al. 2023; Liu et al. 2023; Robinson et al. 2020). Certaines méthodes reposant sur des contraintes plus fortes de correspondance entre voxels ont toutefois été conçues à partir d'images déformées par recalage non linéaire (Fortin et al. 2016; Tian et al. 2022; Torbati et al. 2021).

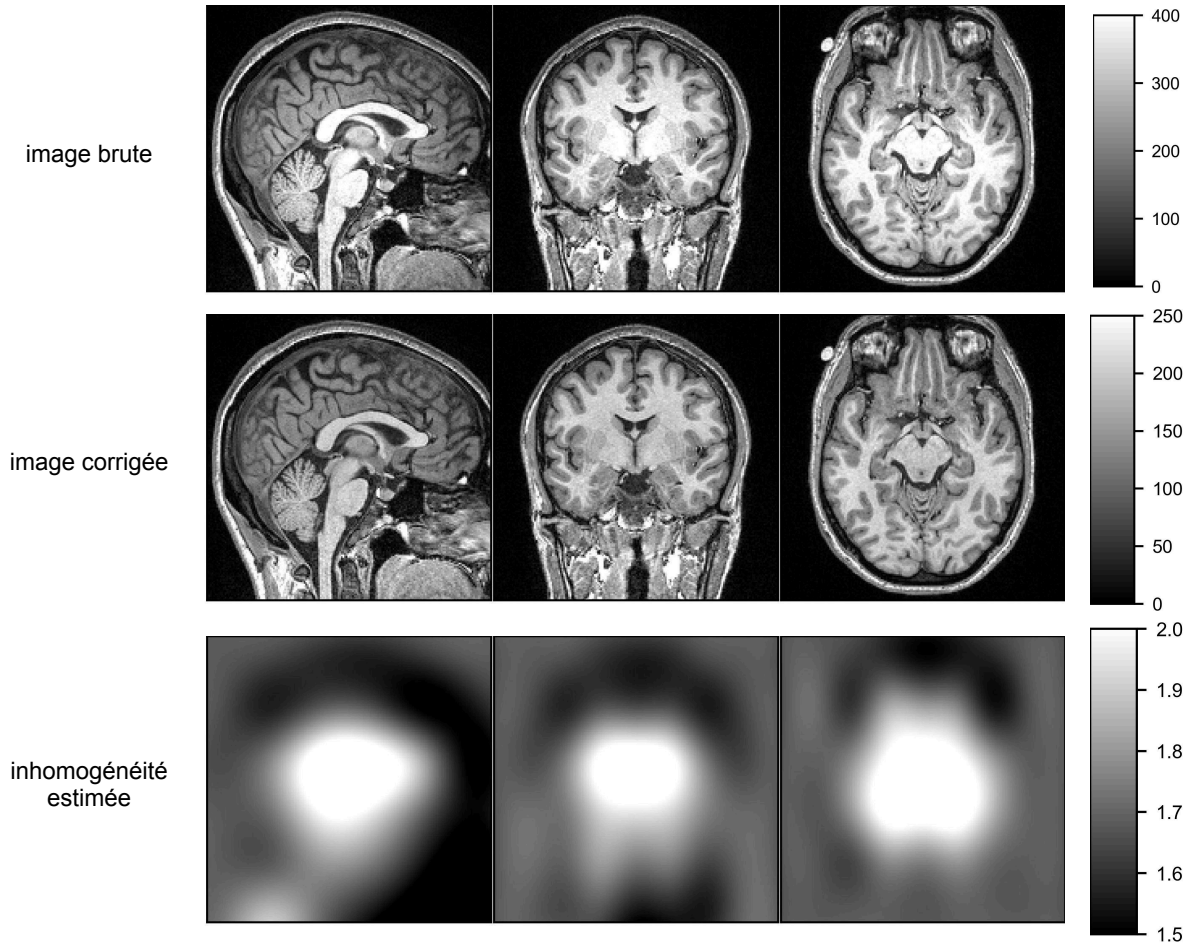


Figure 3 : Exemple d'une correction d'inhomogénéité avec N4ITK. L'inhomogénéité estimée est utilisée pour corriger l'image avec une division par voxel.

2.2.3. Extraction de la boîte crânienne

L'extraction de la boîte crânienne permet de ne conserver que les intensités cérébrales dans l'image (Figure 4). C'est un prétraitement communément réalisé dans les études d'analyse IRM qui permet d'éliminer de l'information inutile et donc de potentiellement réduire la complexité des analyses suivantes sur l'image.

C'est un prétraitement très courant dans le domaine de l'harmonisation qui a été mis en place avec différentes techniques : recalage sur des atlas (e.g. Bashyam et al., 2022 and Pomponio et al., 2020), FSL-BET (Smith 2002) (e.g. Fortin et al. 2016), SPM (e.g. Nguyen et al. 2018), ROBEX (Iglesias et al. 2011) (e.g. Cackowski et al. 2023), HD-BET (Isensee et al. 2019) (e.g. Liu et al. 2023). Néanmoins, certaines études ont fait le choix de garder l'information de la boîte crânienne dans leurs approches d'harmonisation des images (Dewey et al. 2019; Zuo et al. 2021b, 2022).

2.3. Normalisation des intensités

Comme les prétraitements, les techniques de normalisation d'intensité s'appliquent indépendamment sur chaque image. Elles sont particulièrement utiles dans des études multicentriques et sont également souvent appliquées préalablement aux méthodes

d’harmonisation. Leur objectif est de ramener chaque image IRM sur une échelle d'intensité standard. On les distingue des prétraitements car ce sont des méthodes générales de traitement du signal qui sont moins spécifiques à l’IRM.

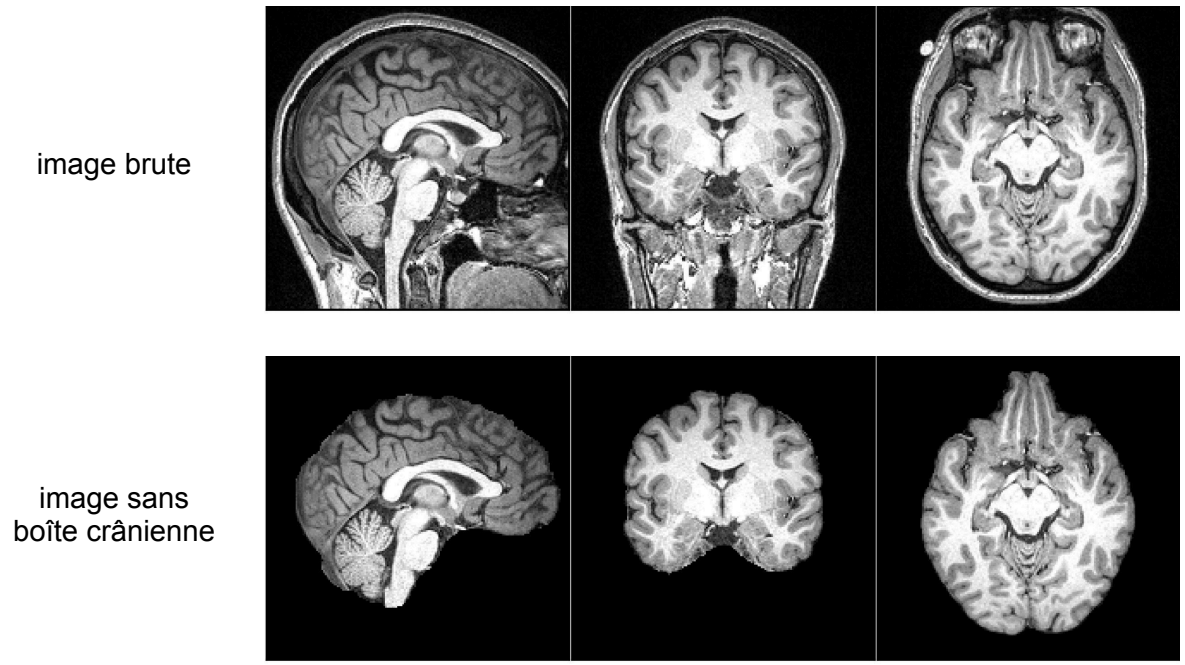


Figure 4 : Exemple d’une extraction de boîte crânienne avec HD-BET.

2.3.1. Normalisation min-max

La normalisation min-max standardise les valeurs minimale et maximale des images IRM : $f(x) = \frac{x - min(x)}{max(x) - min(x)}$. Comme la valeur minimale est presque toujours à 0 dans une image IRM, cela revient à : $f(x) = \frac{x}{max(x)}$. Une hypothèse de cette normalisation est que l’intensité maximale d’une image IRM est un bon indicateur de son échelle d’intensité globale.

Cette normalisation est communément utilisée comme prétraitements à des modèles prédictifs (Gautherot et al. 2021; Iqbal et al. 2019; Li et al. 2020; Mallya et al. 2019; Naser et Deen 2020; Zhao et al. 2018) ou à des méthodes d’harmonisation (Cackowski et al. 2023; Liu et al. 2023; Qu et al. 2020) basés sur de l’apprentissage profond.

2.3.2. Z-score

La normalisation Z-score standardise la moyenne et l’écart-type des intensités : $f(x) = \frac{x - \mu}{\sigma}$ où μ et σ correspondent respectivement à la moyenne et à l’écart-type des intensités de toute l’image ou d’une région spécifique (typiquement le cerveau). Une hypothèse de cette normalisation est que la moyenne et l’écart-type sont des bons indicateurs de l’échelle d’intensité globale.

Certaines études ont utilisé cette normalisation comme préalable à de l’harmonisation (Dinsdale et al. 2021; Palladino et al. 2020; Robinson et al. 2020) ou à de la translation de modalité (Welander et al. 2018). Il est à noter qu’il est possible de calculer la moyenne et

l'écart-type à partir d'un ensemble d'images et non indépendamment pour chaque image (Dar et al. 2019; Nie et al. 2018).

2.3.3. WhiteStripe

La normalisation WhiteStripe est un type particulier de Z-score où la moyenne et l'écart-type sont calculés sur un masque de matière blanche (Shinohara et al. 2014). Sa rapidité et sa robustesse à divers types d'images en entrée en ont fait un outil communément utilisé pour la normalisation d'intensités dans les études multicentriques (Fortin et al. 2016; Gao et al. 2019; Wrobel et al. 2020) et pour le prétraitement à de l'harmonisation (Cackowski et al. 2023; Zuo et al. 2021b).

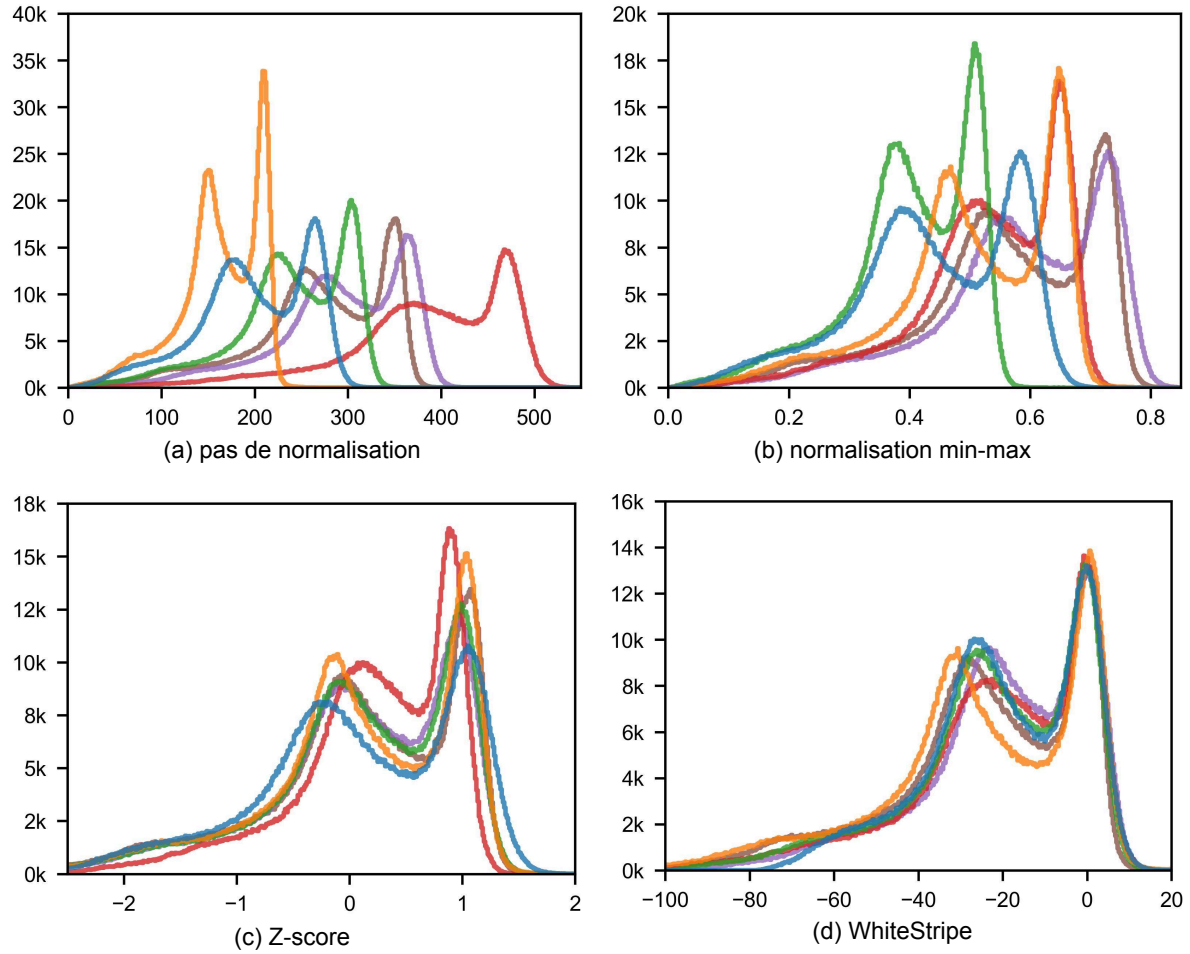


Figure 5 : Histogrammes des intensités cérébrales d'un même sujet avec différentes machines d'acquisition avant et après des normalisations d'intensités. Dans chaque sous-figure, l'axe des X et l'axe des Y indiquent les intensités cérébrales et le nombre de voxels dans la tranche correspondante, respectivement. HD-BET a été utilisé pour la segmentation des cerveaux.

2.3.4. Impact sur les échelles d'intensité

La Figure 5 illustre les distributions d'intensités cérébrales obtenues avant et après les différentes normalisations. Les six images d'un même sujet acquises sur différentes machines sont issues de la base SRPBS (Tanaka et al. 2021).

2.4. L'harmonisation statistique

Les méthodes d'harmonisation statistique ont un niveau de complexité intermédiaire en termes de paramètres estimés. Elles se situent en effet entre les techniques de normalisation d'intensité - qui reposent sur l'estimation d'un ou deux paramètres (section 2.3) au niveau individuel - et les modèles d'apprentissage profond (section 2.5).

2.4.1. Égalisation d'histogrammes

La méthode d'égalisation d'histogrammes (ou histogram matching) initialement proposée par Nyul et al. (2000) et mis à jour par Shah et al. (2011) est un standard pour l'harmonisation des intensités dans des études multicentriques (Bashyam et al. 2022; Fortin et al. 2016; Palladino et al. 2020; Robitaille et al. 2012; Wrobel et al. 2020). Cette méthode repose sur un jeu d'entraînement à partir duquel une échelle d'intensité standard est estimée. Cette échelle d'intensité est calculée comme une moyenne de percentiles d'intensité (percentiles 1,10,20,...,90,99). Après la phase d'entraînement, l'harmonisation se fait par interpolation linéaire par partie, en faisant correspondre l'échelle d'intensité de l'image à harmoniser à l'échelle d'intensité standard. Un masque permet souvent de limiter le calcul des échelles d'intensité à certaines zones d'intérêt (typiquement le cerveau).

Malgré la reconnaissance dont bénéficie cette méthode, certaines études ont mis en avant ses limites, notamment concernant la sensibilité aux artéfacts (e.g. mouvement du patient, homogénéité du champ) et la robustesse à des populations pathologiques présentant des anormalités cérébrales (Fortin et al. 2016; Shinohara et al. 2014). Une érosion de la MG peut par exemple être constatée après une égalisation d'histogrammes (Figure 6).

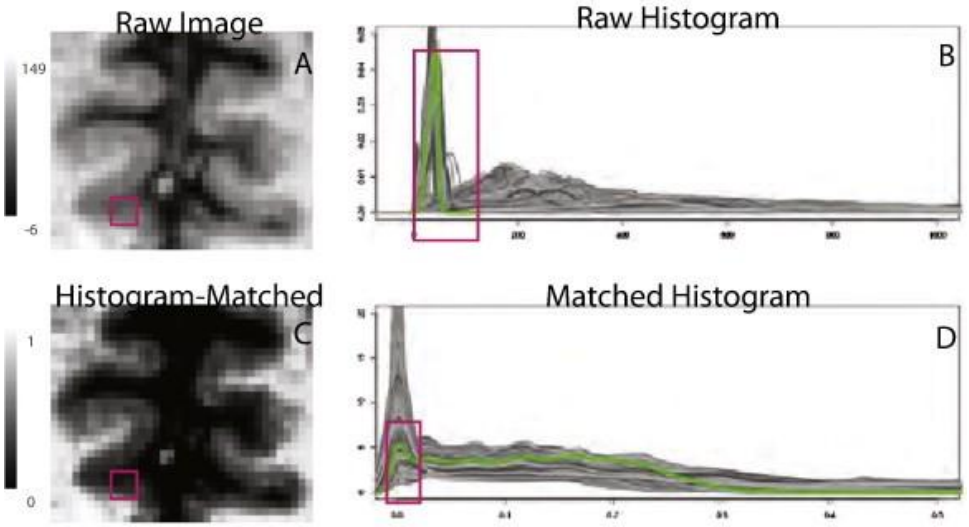


Figure 6 : Illustration de l'échec d'une égalisation d'histogrammes. Première colonne : image IRM de base **(A)** et image après égalisation d'histogrammes **(C)**. Deuxième colonne : histogrammes des images (en vert l'image de la colonne de gauche, en gris des images du même sujet mais sur différentes visites) de base **(B)** et après égalisation d'histogrammes **(D)**. Les carrés rouges indiquent les régions de matière grise disparues après l'égalisation d'histogrammes. Figure adaptée de Shinohara et al. (2014).

2.4.2. RAVEL

La méthode RAVEL proposée par Fortin et al. (2016) est une extension de la normalisation WhiteStripe (section 2.3.3). Les auteurs mettent en avant que WhiteStripe standardise les intensités de MB mais ne permet pas une bonne homogénéisation de la MG.

RAVEL est un modèle linéaire estimant l'effet de la variabilité biologique et l'effet de facteurs *inconnus* sur les intensités cérébrales. L'objectif est de supprimer la variabilité due aux facteurs inconnus. L'effet de la variabilité biologique est estimée grâce à des covariables fournies au modèle (e.g. statut pathologique, âge, sex). Les covariables relatives aux facteurs inconnus sont estimées à partir du postulat que les intensités de LCS ne sont pas associées au statut pathologique ou à d'autres covariables biologiques. Après une segmentation automatique du cerveau, le masque de LCS est alors considéré comme une région contrôle et une décomposition en valeurs singulières est utilisée pour estimer les facteurs inconnus.

Les auteurs mettent en avant la capacité de RAVEL à harmoniser la MG, la MB et le LCS (Figure 7).

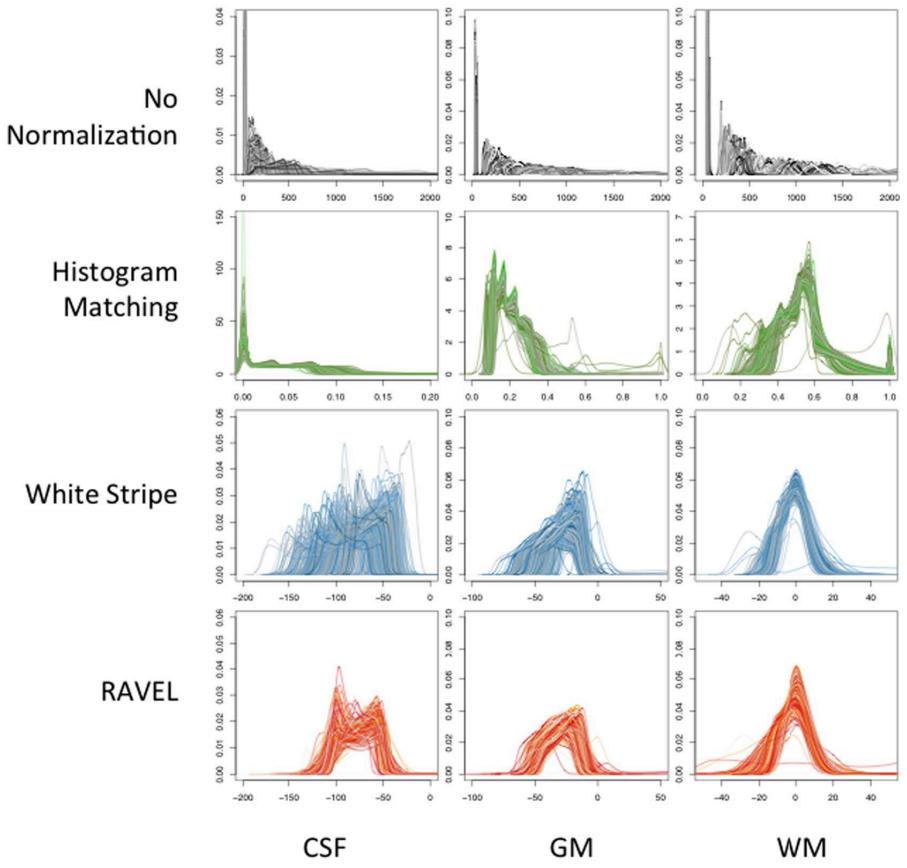


Figure 7 : Histogrammes des intensités par tissu cérébral avant et après normalisation. Chaque courbe représente l'histogramme d'intensité pour un sujet. CSF : liquide cérébrospinal; GM: matière grise; WM: matière blanche. Figure de Fortin et al. (2016)**.**

La principale originalité de la méthode RAVEL est l'intégration d'une connaissance à priori sur la variabilité dans le LCS. Les auteurs montrent en outre de moins bons résultats sans cet à priori et en fournissant directement des variables techniques au modèle (site d'acquisition, constructeur du scanner et intensité du champ). Un autre point

méthodologique important est que RAVEL ne vise la conservation que des variabilités associées aux covariables biologiques fournies pour l'entraînement et l'inférence.

2.4.3. ComBat

Contrairement aux méthodes vues précédemment, la méthode ComBat n'est pas conçue pour le traitement d'images mais plutôt pour des données de moindre dimension comme des caractéristiques extraites des images. Introduite par Johnson et al. (2007) pour limiter la variabilité technique dans le contexte de l'expression de gènes, elle a ensuite été adaptée en IRM pour l'harmonisation inter-sites de cartes d'anisotropie fractionnelle (AF) et de diffusivité moyenne (DM) (Fortin et al. 2017) et de mesures d'épaisseurs corticales (Fortin et al. 2018).

Le modèle ComBat est construit de manière similaire à RAVEL avec une prise en compte de covariables biologiques dont les effets sont destinés à être conservés. En revanche, pour la variabilité technique, ComBat n'utilise pas de région contrôle comme RAVEL mais sépare les données par site (ou avec une autre covariable dont on cherche à supprimer les effets). Il est à noter qu'aucune des méthodes vues précédemment ne prend en compte cette information. Une conséquence importante est que ComBat s'applique sur des données regroupées (e.g. par site) et non sur des données au niveau individuel. Maikusa et al. (2021) ont essayé de déterminer le nombre minimal de sujets nécessaires pour l'utilisation de ComBat. Deux autres caractéristiques méthodologiques sont importantes. La première est que les effets de site sont modélisés par un facteur additif et un facteur multiplicatif. L'objectif associé est la standardisation de la moyenne et de la variance des caractéristiques à travers les sites. La deuxième est l'adoption d'une approche bayésienne pour l'estimation des paramètres du modèle. Cette approche intègre un *a priori* liant les effets de site sur les différentes caractéristiques à harmoniser. Le modèle n'est donc pas appliqué indépendamment sur chaque caractéristique et toute l'information est exploitée pour l'harmonisation de chacune.

ComBat est devenu un standard pour l'harmonisation dans des études multicentriques. Radua et al. (2020) l'ont par exemple utilisé pour harmoniser un large jeu de données d'épaisseurs corticales issu du consortium ENIGMA et ont montré un renforcement de biomarqueurs séparant des sujets sains de sujets de sujets schizophrènes. Yu et al. (2018) ont montré que ComBat pouvait permettre des analyses plus fiables et plus efficaces sur des mesures de connectivité fonctionnelle.

Différentes extensions de la méthode ont été développées ces dernières années. Les principales sont les suivantes :

- Pomponio et al. (2020) ont proposé une modélisation non-linéaire de l'évolution de volumes corticaux et sous-corticaux avec l'âge. En utilisant un modèle additif généralisé, l'âge est utilisé comme prédicteur non-linéaire et le modèle est optimisé par maximum de vraisemblance. Les auteurs prétendent que leur modèle capture mieux l'évolution des volumes avec l'âge que des modèles linéaires ou quadratiques.
- Beer et al. (2020) ont adapté le modèle aux études longitudinales pour une prise en compte explicite de la corrélation des mesures pour un même sujet. L'hypothèse est que l'exploitation de l'information longitudinale aide le modèle à distinguer les variabilités liées au site des variabilités biologiques.
- Maikusa et al. (2021) ont évalué ComBat sur un jeu de données de sujets "voyageurs" (i.e. sujets ayant été scannés sur différents sites) et l'ont comparée à une approche ComBat adaptée pour de telles données. L'approche en question

n'intègre pas de covariables biologiques à conserver et utilise uniquement les sujets voyageurs pour préserver l'information pertinente. Les expériences menées montrent un avantage de cette dernière méthode, ce qui suggère que les covariables biologiques ne sont pas suffisantes pour la préservation de l'information individuelle.

- Chen et al. (2022) ont étendu ComBat pour harmoniser les covariances des caractéristiques IRM (épaisseurs corticales dans l'étude), en plus des moyennes et des écart-types de l'approche originale.
- Wachinger et al. (2021) ont ajouté au modèle ComBat un vecteur collectant des covariables relatives à l'acquisition (constructeur et intensité de champ dans les expériences menées) dont les effets doivent également être annulés. En plus de ces covariables explicitement spécifiées, le modèle estime avec une Analyse en Composantes Principales (ACP) des covariables inconnues pouvant induire des variabilités techniques dans les données. L'hypothèse est que l'information du site ne recouvre pas toutes les variabilités techniques potentielles et que d'autres informations connues ou inconnues peuvent faciliter la capture de ces variabilités non souhaitées dans les données.

Malgré les nombreuses utilisations de méthodes de type ComBat dans des études multicentriques, Richter et al. (2022) ont montré des limites dans son utilisation sur des données structurelles. Dans leur étude sur des sujets voyageurs, peu de variabilités inter-sites ont été constatées sur des mesures d'épaisseurs corticales et de volumétrie. Après l'application de méthodes ComBat (Beer et al. 2020; Fortin et al. 2017; Pomponio et al. 2020), certaines de ces variabilités ont été accentuées et certains motifs biologiques obscurcis. En revanche, d'autres résultats montrent l'intérêt de cette méthode sur des mesures de diffusion (AF et DM).

Une autre limite de ces méthodes est qu'elles ne sont pas destinées à l'harmonisation d'images IRM complètes. Elles requièrent alors une harmonisation spécifique pour chaque ensemble de caractéristiques. En outre, Zuo et al. (2021b) ont décrit ces méthodes comme des post-traitements destinés à corriger les erreurs d'un outil de segmentation liées à des variabilités inter-sites. Selon eux, les outils statistiques comme ComBat ne sont pas capables de corriger ces variabilités si elles sont trop importantes.

2.4.4. NeuroHarmony

Garcia-Dias et al. (2020) ont proposé l'outil NeuroHarmony pour contourner une limite importante de la méthode ComBat (et de beaucoup de méthodes d'harmonisation) qui est que les données à harmoniser doivent venir d'un site connu et vu pendant l'entraînement du modèle. Leur idée est d'apprendre à reproduire les corrections de la méthode ComBat à partir des informations d'une image (sans utiliser l'information du site). Ils ont pour cela tout d'abord harmonisé un large jeu de données multicentriques de mesures volumétriques avec ComBat. Ils ont ensuite utilisé un modèle supervisé, une forêt d'arbres décisionnels, pour apprendre à prédire une correction de volume ComBat à partir du volume original, de métriques de qualités d'image (IQMs, Esteban et al. 2017), de l'âge et du sexe. La structure du modèle est illustrée dans la Figure 8.

Bien que cette méthode reste moins populaire que ComBat, son originalité est notable car c'est la première approche d'harmonisation permettant de traiter des données provenant de sites inconnus (si l'on excepte les techniques de normalisation). Comme nous le verrons avec certaines approches d'apprentissage profond (section 2.5), une idée clé pour cela est

que l'information du site n'est utilisée que pendant l'entraînement (à travers ComBat dans NeuroHarmony).

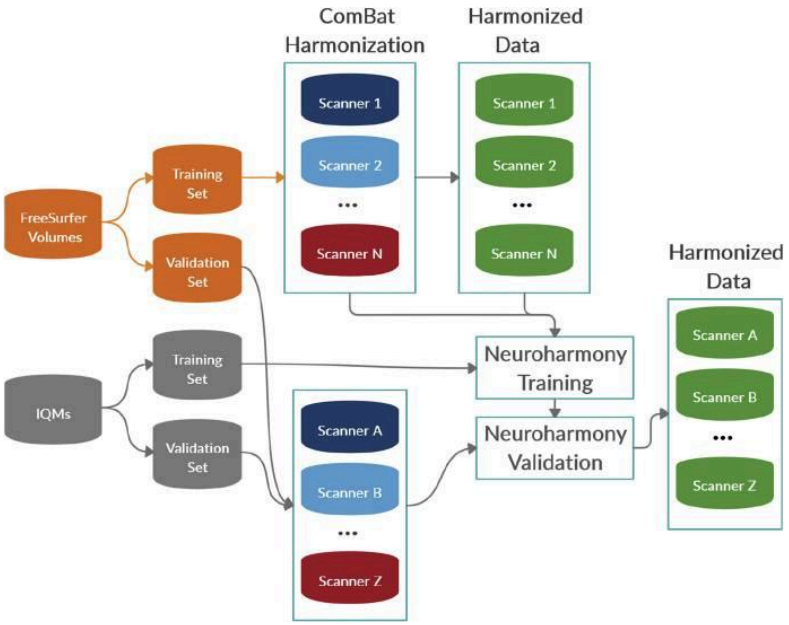


Figure 8 : Illustration du modèle NeuroHarmony. IQMs correspond aux métriques de qualité d'image utilisées pour prédire les corrections ComBat. Figure de Garcias-Dias et al. (2020).

2.5. Harmonisation par apprentissage profond

Nous nous intéressons dans cette section à différentes approches d'apprentissage profond qui ont été utilisées pour traiter directement les images IRM au niveau des voxels. Ces méthodes ont une complexité bien supérieure à celles vues jusqu'ici ; elles impliquent en effet très souvent l'optimisation de plusieurs millions de paramètres à travers l'entraînement de réseaux de neurones.

2.5.1. Harmonisation orientée prédictions

L'optimisation des performances de modèles de prédiction est l'une des motivations aux travaux de recherche sur l'harmonisation. Des méthodes d'harmonisation orientées spécifiquement pour des tâches de prédiction ont alors été mises en place. Dinsdale et al. (2021) ont par exemple développé un modèle de prédiction basé sur de l'apprentissage supervisé en mettant en place trois sous-modèles : un extracteur de caractéristiques, un prédicteur de labels (i.e. la/les variable(s) biologique(s) à prédire) et un classifieur de domaines. La procédure d'entraînement est divisée en trois étapes : (i) optimisation de l'extracteur de caractéristiques et du prédicteur de labels pour optimiser les performances de prédiction, (ii) optimisation du classifieur de domaines pour l'identification du domaine d'origine (e.g. le site) à partir des caractéristiques extraites et (iii) optimisation de l'extracteur de caractéristiques pour confondre le classifieur de domaine. L'objectif est alors de construire un espace latent contenant l'information biologique d'intérêt mais qui ne permet pas d'identifier le site d'acquisition des images IRM. Guan et al. (2021) ont mis en place une approche similaire qui utilise un mécanisme d'attention en vue de se focaliser sur les régions pertinentes pour la prédiction de la variable biologique d'intérêt.

Wang et al. (2022) intègre également un extracteur de caractéristiques et un prédicteur de labels dans leur modèle. L'invariance au site n'y est cependant pas produite par un entraînement antagoniste avec un discriminateur mais par une succession de *fine-tuning*. L'originalité de leur méthodes vient aussi des tâches de prédictions *auxiliaires* qu'ils intègrent pendant l'entraînement. Leur idée est que dans le cas où l'on ne dispose pas du label d'intérêt associé aux images IRM d'un site, on peut exploiter d'autres labels souvent disponibles comme l'âge et le sexe qui peuvent être liés au label principal. La structure du modèle est illustrée dans la Figure 9.

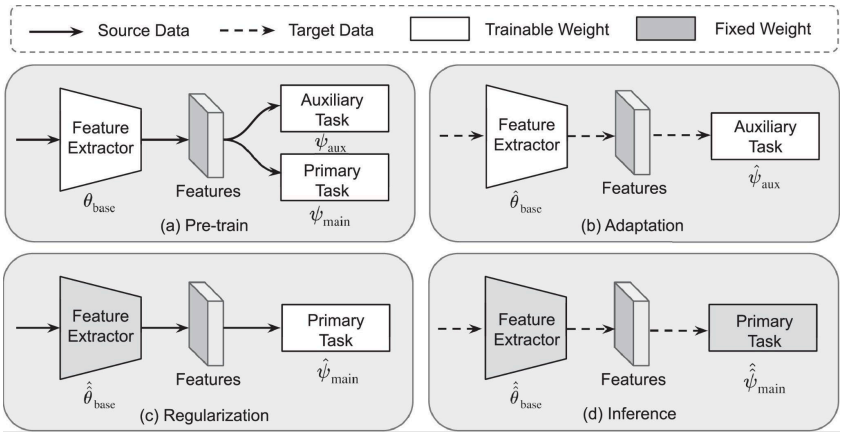


Figure 9 : Illustration d'un modèle d'harmonisation orientée prédictions. Figure de Wang et al. (2022).

Ces approches d'harmonisation orientée prédictions peuvent être pertinentes si l'on souhaite se focaliser sur une tâche spécifique. Elles requièrent cependant un nouvel entraînement complet pour chaque nouvelle tâche et sont de plus limitées au contexte de la prédiction supervisée.

2.5.2. Modèles génératifs supervisés

Dewey et al. (2019) ont proposé l'un des premiers modèles génératifs pour l'harmonisation inter-sites. Il consiste en un générateur avec une architecture U-net entraîné de manière supervisée à harmoniser des images de différents domaines. La supervision repose sur des sujets voyageurs ayant été scannés avec les deux protocoles de l'étude. Une fonction de coût à l'échelle du voxel guide l'entraînement du générateur. Dans l'étude, les auteurs n'ont utilisé que 10 sujets pour l'entraînement, suggérant que leur approche ne requiert pas de base de données très grande. Une autre particularité est l'harmonisation simultanée de quatre contrastes IRM.

Qu et al. (2020) ont mis en place WATNet, un modèle similaire pour harmoniser les images d'un scanner 3 Tesla (T) vers un scanner 7T. Une différence notable est l'intégration d'ondelettes qui a pour but de mieux capturer les différentes fréquences de motif dans l'image qui vont des contrastes de tissus de basses fréquences aux détails anatomiques de haute fréquence. La structure du modèle est illustrée dans la Figure 10.

Tian et al. (2022) ont développé un modèle d'harmonisation de cartes de volumes de MG basé sur du démêlage de l'information du site et de l'information anatomique dans l'image IRM. Pour chaque paire de sites, 4 encodeurs (un encodeur d'anatomie et un encodeur de contraste pour le site *source* et pour le site *cible*) et 2 décodeurs sont établis (un pour le site

source et pour le site *cible*). Différentes fonctions de coûts relatives aux espaces et aux images générées guident l'entraînement des encodeurs et des décodeurs en utilisant les sujets voyageurs de l'étude.

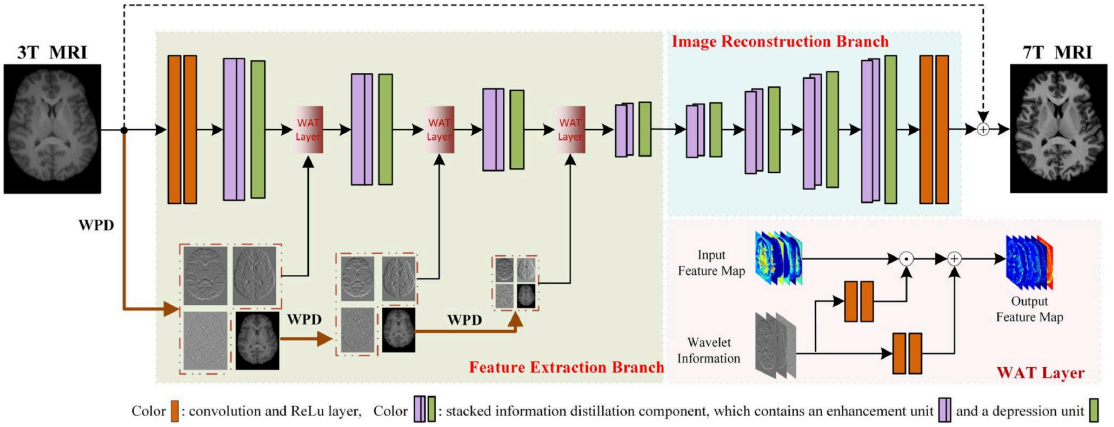


Figure 10 : Illustration de WATNet pour l'harmonisation d'un scanner 3 Tesla vers un 7 Tesla. Figure de Qu et al. (2020).

Torbati et al. (2021) ont proposé une approche plus simple avec pour chaque scanner de l'étude, un encodeur produisant des *images latentes* et un décodeur produisant une combinaison linéaires de ces images. Les différentes fonctions de coûts encouragent la similarité des images latentes et des images décodées sur les différents scanners pour un même sujet ainsi que la préservation de l'information de l'image originale à travers une fonction de coût de reconstruction. L'une des différences notables de cette méthode par rapport aux trois précédentes est qu'elle peut apprendre à harmoniser plus de deux sites *simultanément*.

Ces approches sont pertinentes quand on souhaite harmoniser des domaines pour lesquels on dispose de sujets voyageurs. Elles ont en effet l'avantage d'être efficaces avec relativement peu d'images IRM d'entraînement. En revanche, elles ne sont pas applicables sans sujets voyageurs, ce qui est une importante limitation pour l'harmonisation de données multicentriques.

2.5.3. Modèles génératifs non-supervisés

2.5.3.1. Approches CycleGAN

CycleGAN (Zhu et al. 2017) est une approche de référence pour des tâches de translation d'images. L'objectif est d'apprendre une transformation faisant correspondre deux ensembles d'images. Soient deux domaines X et Y, le modèle doit transformer une image de X (respectivement Y) en une image ressemblant à une image de Y (respectivement X) tout en conservant les informations *individuelles* de l'image originale. Dans CycleGAN, l'information individuelle à conserver correspond à l'information qui ne permet pas de distinguer une image du domaine X d'une image du domaine Y.

CycleGAN est implémenté avec des réseaux antagonistes génératifs (GAN, Goodfellow et al. 2014) qui permettent de générer des images réalistes dans les domaines X et Y (2 générateurs et 2 discriminateurs). L'entraînement est contraint par une fonction de coût de *consistance du cycle* encourageant le modèle à reproduire l'image originale lorsqu'une image est transférée de X vers Y puis de Y vers X (ou inversement). La structure du modèle

est illustrée dans la Figure 11. L'intérêt de la méthode est qu'elle ne requiert pas d'appariement entre les deux domaines pour l'entraînement mais simplement des images représentatives pour chacun.

De nombreuses études ont expérimenté CycleGAN pour l'harmonisation inter-sites d'images cérébrales T1w (Chen et al. 2021; Enriquez Calzada 2021; Gebre et al. 2023; Nguyen et al. 2018; Palladino et al. 2020). Zhong et al. (2020) ont harmonisé des métriques de diffusion avec DualGAN (Yi et al. 2017), un modèle très similaire à CycleGAN qui diffère par l'architecture des générateurs et la fonction de coût antagoniste. Ces approches non-supervisées ont aussi été beaucoup appliquées à la translation de séquences IRM ou de modalité (Dar et al. 2019; Welander et al. 2018; Wolterink et al. 2017). Dans leur étude de synthèse de tomodensitométrie à partir d'images IRM, Wolterink et al. (2017) compare CycleGAN avec pix2pix - une approche similaire utilisant un appariement entre les deux domaines d'image (Isola et al. 2017) - et montrent une supériorité de CycleGAN même dans le cas où l'on dispose de données appariées. Ils expliquent cela par l'impossibilité d'avoir un alignement parfait entre les différentes modalités, ce qui biaise l'apprentissage supervisé à l'échelle du voxel. Dans leurs travaux sur de la translation de domaines entre images T1w et T2-pondérées (T2w), Welander et al. (2018) tirent des conclusions différentes puisque leurs expériences montrent que, bien que les images générées par CycleGAN soient réalistes, pix2pix et un générateur simplement supervisé donnent de meilleurs résultats avec des métriques quantitatives.

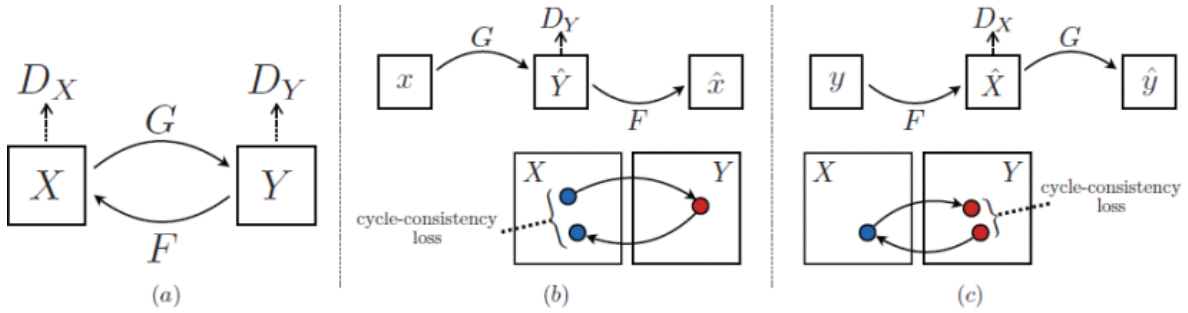


Figure 11 : Illustration du modèle CycleGAN. (a) Le modèle contient deux fonctions de translation $G : X \rightarrow Y$ et $F : Y \rightarrow X$, et les discriminateurs correspondant D_Y et D_X. D_Y encourage G à générer des images indistinguables de celles de Y et vice versa pour D_X, F et X. La fonction de coût de consistance du cycle (cycle-consistency loss) encourage les deux générateurs à préserver les informations non liées aux domaines dans l'image originale : **(b)** $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$ et **(c)** $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. Figure de Zhu et al. (2017).

CycleGAN ayant été développé originellement pour des images au format RVB, des chercheurs ont apporté des modifications au modèle pour une application plus adaptée aux images IRM. Hognon et al. (2019) ont par exemple mis en place un apprentissage en deux étapes mêlant CycleGAN et pix2pix. CycleGAN est d'abord entraîné brièvement à harmoniser un ensemble d'images issues d'un jeu de données multicentriques et une image *template* de sujet sain *data-augmentée*. Le jeu de données multicentrique est ensuite harmonisé vers le template avec CycleGAN. Une deuxième étape utilise ensuite pix2pix pour harmoniser les images originales vers le domaine des images harmonisées préalablement par CycleGAN. L'idée des auteurs est que la première étape sert de *data-augmentation* pour générer de multiples images dans le style du template de référence. La brièveté de l'entraînement de CycleGAN vise à éviter l'altération des structures

cérébrales. Ces structures seraient ainsi mieux préservées durant l'entraînement supervisé suivant. Avec le même objectif, Chang et al. (2022) ont également mis en place un entraînement en appliquant une technique d'égalisation d'histogrammes après une translation CycleGAN. D'autres chercheurs ont plutôt opté pour de nouvelles fonctions de coûts afin de mieux préserver l'information anatomique. Kieselmann et al. (2021) ont par exemple ajouté une pénalisation des différences entre les masques cérébraux des images d'entrée et ceux des images générées, sans toutefois préciser comment sont calculés ces masques. Modanwal et al. (2021) ont eux ajouté une fonction de coût pénalisant la perte d'*information mutuelle* afin de mieux préserver la forme des poitrines et les caractéristiques des tissus dans leurs images IRM mammaires. Avec la même idée, Xiang et al. (2018) ont intégré une version modifiée du SSIM - une métrique de quantification de similarités structurelles (Wang et al. 2004) - pour l'entraînement à la préservation de l'information anatomique des images.

Ces adaptations conservent le principe général de CycleGAN. Certaines modifications méthodologiques plus innovantes ont également été expérimentées. Sinha et al. (2021) ont par exemple testé l'intégration de *masques d'attention* (Tang et al. 2019) pour que leur modèle d'harmonisation inter-sites se focalise sur les zones les plus discriminatives entre les sites. Yan et al. (2019) ont de leur côté ajouté une fonction de coût basée sur des caractéristiques extraites par un modèle de segmentation, avec l'hypothèse qu'un modèle de segmentation bien entraîné extrait des caractéristiques pertinentes des images. Ils ont de plus remplacé la fonction de consistance du cycle originale par le SSIM pour mieux préserver les structures anatomiques. Enfin, Qin et al. (2022) ont combiné CycleGAN avec du transfert de style pour de la translation entre modalités IRM : les générateurs prennent en entrée une image du domaine cible en plus de l'image à harmoniser. Les caractéristiques de contenu et de style sont fusionnées par *normalisation d'instance adaptative* (Huang et Belongie 2017). Les auteurs mettent en avant que l'image synthétisée et l'image cible ont ainsi des styles "plus identiques à différentes échelles". Une autre particularité de leur méthode est l'ajout de bruits aléatoires à différents niveaux des générateurs afin de rendre le modèle plus robuste aux bruits liés à l'acquisition des images.

2.5.3.2. Translation d'images multi-domaine

Un défaut de CycleGAN est la nécessité d'entraîner un modèle pour chaque paire de domaines de l'étude. Si l'on souhaite par exemple harmoniser des images provenant de N sites d'acquisition, N-1 CycleGAN doivent être entraînés pour translater les images vers un site de référence. Les N-1 modèles étant de plus indépendants, chacun d'entre eux ne bénéficie pas de l'entièreté du jeu de données à disposition. Des modèles de translation d'image autorisant l'inclusion d'un nombre arbitraire de domaines ont alors été développés pour pallier cela.

StarGAN, une extension de CycleGAN qui permet d'apprendre des translations entre plus de deux domaines d'image avec simplement un générateur et un discriminateur (Choi et al. 2018), en est un exemple notable. L'idée principale de StarGAN est que le générateur n'est pas simplement conditionné par l'image d'entrée, mais également par le domaine cible (Figure 12). Fuhai et Tang (2021) ont appliqué ce modèle pour translater des images entre différentes modalités IRM.

Le modèle StarGAN v2 a ensuite été proposé par Choi et al. (2020). Les auteurs mettent en avant une limitation de la première version qui est la génération déterministe pour chaque domaine, le générateur étant conditionné simplement par le domaine cible en plus de

l'image d'entrée. StarGAN v2 produit des codes de style spécifiques à chaque image pour chaque domaine et utilise ces codes de style pour conditionner la génération d'images, ce qui permet une multitude de styles pour chaque domaine. Bashyam et al. (2022) ont utilisé cette approche pour harmoniser un large jeu de données contenant des images IRM provenant de six scanners différents.

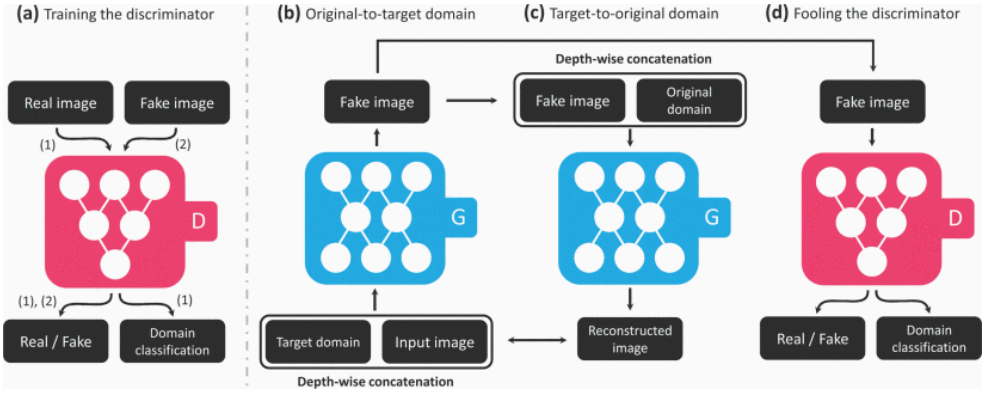


Figure 12 : Illustration du modèle StarGAN. (a) Le discriminateur (D) apprend à distinguer les fausses images des images réelles et à classifier le domaine des images réelles. **(b)** Le générateur (G) génère une fausse image à partir d'une image d'entrée et d'un domaine cible. **(c)** G essaie de reconstruire l'image originale à partir de l'image fausse et du domaine de l'image originale. **(d)** G essaie de générer une image indistinguable des images réelles et classifiée dans le domaine cible par D. Figure de Choi et al. (2018).

Gao et al. (2019) ont également développé un modèle intégrant des translations entre de multiples domaines pour l'harmonisation inter-sites d'images T2-FLAIR. Un générateur *universel* est entraîné à transformer les images des différents sites vers un site de référence. En parallèle, des générateurs sont entraînés à transformer les images du site de référence vers chaque site (un générateur pour chaque site). Comme StarGAN et StarGAN v2, le modèle est entraîné avec des réseaux antagonistes génératifs et des contraintes de consistance du cycle de manière unifiée (un seul modèle peu importe le nombre de domaines). En revanche, si la base d'entraînement est suffisamment diversifiée, il permet de plus d'harmoniser des images provenant de domaines non vus pendant l'entraînement, ce qui peut être particulièrement souhaitable pour l'harmonisation inter-sites d'images IRM.

Fatania et al. (2022) ont mis en place un auto-encodeur (HarMOnAE) pour une harmonisation multisite applicable à n'importe quelle image après l'entraînement. L'approche est plus simple que celle de Gao et al. et repose simplement sur un espace latent rendu "agnostique" au site par un entraînement antagoniste avec un discriminateur chargé d'identifier le site d'origine. Le décodeur est conditionné par le label du site vers lequel on souhaite *transférer* l'image d'entrée. L'architecture du modèle est illustrée dans la Figure 13.

2.5.3.3. Transfert de style

En apprentissage profond, le transfert de style peut avoir une définition relativement large englobant toutes les approches de translations d'image à image. Dans ce manuscrit, nous distinguons néanmoins les méthodes de translation de domaine (e.g. CycleGAN) des méthodes de transfert de style. Ces dernières appliquent un style d'une seule image à une autre image. Une information de domaine peut éventuellement être incluse pendant l'entraînement, mais pas à l'inférence. Il s'ensuit que ces approches peuvent être appliquées

à n'importe quelle image après l'entraînement, peu importe le domaine (au moins théoriquement). Il doit être noté que StarGAN v2 (Choi et al. 2020), malgré son utilisation de style spécifique à une image, ne rentre pas vraiment dans cette catégorie étant donné que les codes de style générés sont affiliés à un domaine vu pendant l'entraînement.

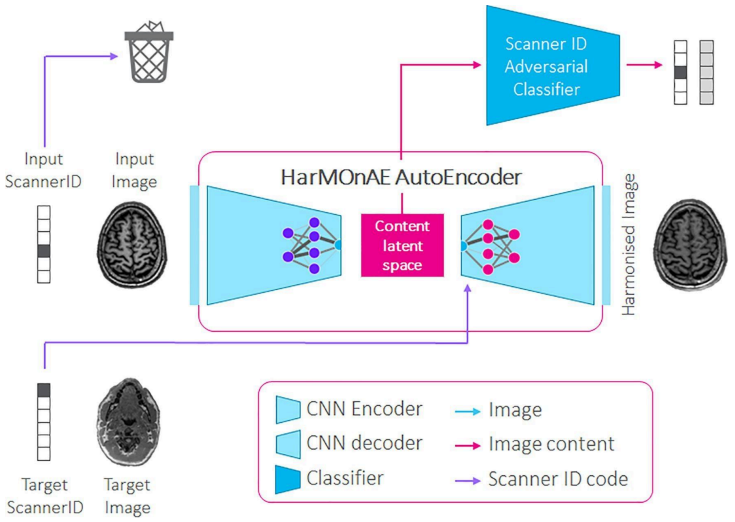


Figure 13 : Illustration du modèle HarMOnAE. L'espace latent rendu "agnostique" au site et le décodeur conditionné par le label du site cible permettent l'harmonisation. Figure de Fatania et al. (2022).

Le transfert de style a beaucoup été utilisé pour l'harmonisation inter-sites ces dernières années. Zuo et al. (2021b) ont par exemple mis en place un modèle, CALAMITI, incluant un encodeur d'anatomie, un encodeur de contraste et un décodeur permettant de synthétiser une image IRM à partir de ces deux informations. Afin de démêler les informations d'anatomie et de contraste, CALAMITI utilise une image T1w et une image T2w pour chaque sujet de la base d'entraînement. Le modèle s'entraîne de manière supervisée à transformer les images T1w en T2w et inversement. Un discriminateur est intégré au modèle pour encourager la consistance des informations anatomiques entre les différents sites de l'étude. Après l'entraînement, le modèle est utilisé pour de l'harmonisation inter-sites. CALAMITI repose ainsi sur deux hypothèses fortes. La première est que les images cérébrales T1w et T2w contiennent la même information anatomique. La deuxième est qu'en apprenant la capture des différences de contraste entre les deux modalités, le modèle sera capable d'éliminer les variabilités inter-sites. Bien que le modèle entraîné à démêler contraste et anatomie puisse à priori être appliqué à n'importe quelle image, les auteurs proposent une procédure de fine-tuning pour l'adaptation à de nouveaux sites.

Cackowski et al. (2023) se sont inspirés de CALAMITI pour mettre en place ImUnity, un auto-encodeur d'harmonisation d'images cérébrales T1w. ImUnity suit également un apprentissage supervisé. En revanche, il ne requiert pas de paires d'images T1w et T2w pour l'entraînement mais repose sur des modifications de contrastes consistant en des transformations *gamma*. L'encodeur du modèle est de plus entraîné de manière antagoniste à un module prédisant le site d'origine à partir de l'espace latent. Un module de préservation d'informations liées à des variables biologiques peut optionnellement être également connecté à l'espace latent. Enfin, un entraînement antagoniste au niveau des images encourage l'auto-encodeur à produire des images réalistes. Deux hypothèses sont que les

fonctions de transformations gamma préservent l'information anatomique et qu'elles simulent efficacement des effets de site.

Le modèle d'harmonisation d'images IRM mammaires proposé par Cao et al. (2023), StyleMapper, présente des similarités avec CALAMITI et ImUnity. Comme dans CALAMITI, deux encodeurs sont chargés de démêler contraste et anatomie et comme dans ImUnity, des modifications d'images sont produites pour la mise en place d'un entraînement supervisé (Figure 14). En revanche, StyleMapper se concentre sur l'harmonisation de contraste à l'échelle des images et ne prend pas en compte l'information du site.

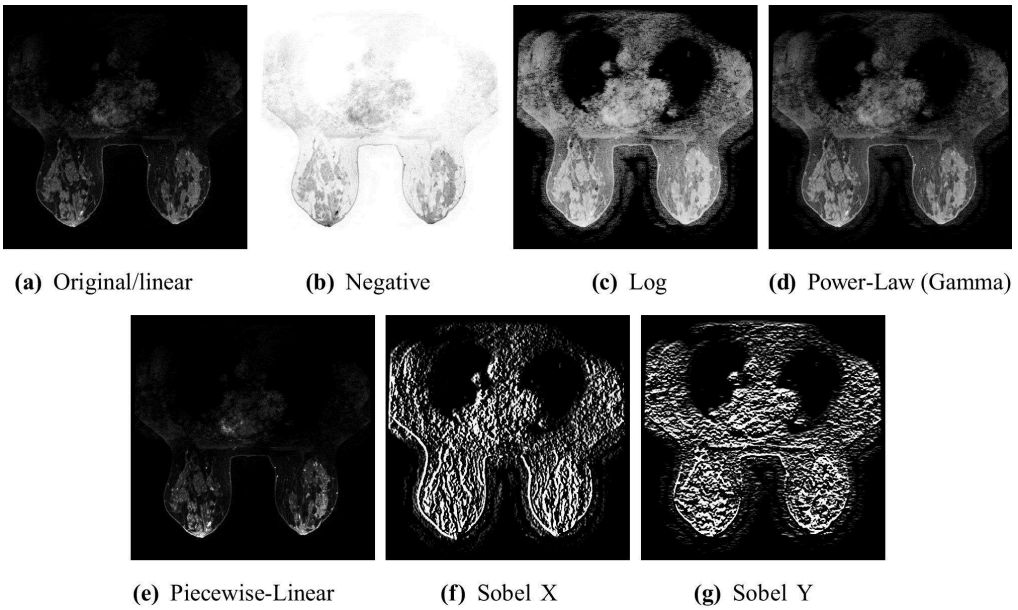


Figure 14 : Exemples de transformations d'une image IRM pendant l'entraînement de StyleMapper. Figure de Cao et al. (2023).

Zuo et al. (2022) ont développé une approche plus simple pour séparer les informations anatomiques et de contraste. Elle repose sur le postulat que deux coupes d'une même image IRM avec différentes orientations partagent le même contraste mais diffèrent anatomiquement (postulat semblable également utilisé dans CALAMITI et ImUnity). À partir de deux telles coupes x et x', trois fonctions de coût encouragent (i) la capacité du modèle à reconstruire x à partir de son code anatomique et du code de contraste de x', (ii) une *information mutuelle* maximale entre x et le code de contraste de x' et (iii) une information mutuelle minimale entre le code anatomique de x et le code de contraste de x'. Les différents modules et les fonctions de coût sont illustrés dans la Figure 15.

Liu et al. (2023) se sont inspirés de StarGAN v2 (Choi et al. 2020) pour leur modèle d'harmonisation. La différence notable est qu'aucune information de domaine n'est fournie au modèle. Selon les auteurs, ce choix est pertinent en IRM étant donné que des variabilités techniques proviennent d'autres facteurs que le site ou le scanner, par exemple les mises à jour logicielles. On peut cependant remarquer que les auteurs n'expliquent pas comment le modèle est censé distinguer les variabilités biologiques des variabilités techniques.

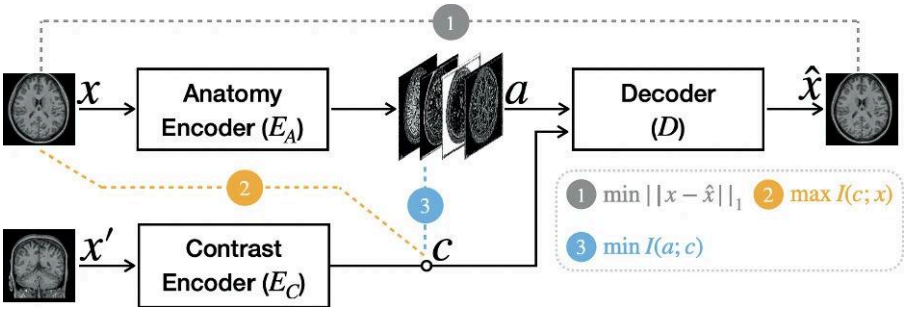


Figure 15 : Illustration d'un modèle harmonisation basé sur démêlage des informations anatomiques et de contraste. I(.; .) dénote l'information mutuelle. Figure de Zuo et al. (2022).

2.5.3.4. Harmonisations 2D/3D

Les modèles de translation d'image ont été développés à l'origine pour traiter des images au format RVB. Leurs traitements sont alors effectués en 2D, très souvent avec des convolutions 2D. En IRM, les images sont souvent des volumes mais une grande majorité des modèles génératifs qui y sont utilisés sont restés 2D, notamment parmi les non-supervisés. La perte résultante d'informations spatiales n'a jamais été justifiée ou a été dictée par des contraintes matérielles ou pratiques (Dewey et al. 2020; Liu et al. 2023; Nguyen et al. 2018; Palladino et al. 2020; Zuo et al. 2021b). Certains chercheurs ont mis en avant la rapidité de calcul et le moindre besoin en données pour justifier leur modèle 2D (Cackowski et al. 2023; Dewey et al. 2019). Le premier argument est difficile à contester puisqu'il est effectivement attendu que des modèles moins complexes soient plus rapides à entraîner. Le deuxième argument est plus discutable et serait difficile à valider ou invalider de manière catégorique dans le cas général.

Malgré la prépondérance des approches 2D, certaines études ont mis en place des procédures spécifiques à la synthèse d'images 3D. Une technique consiste par exemple à entraîner trois modèles d'harmonisation respectivement dans les trois orientations orthogonales (axial, coronal et sagittal) et à l'inférence, un volume pour chaque orientation est généré avant que les trois volumes ne soient combinés par une médiane par voxel pour obtenir l'image IRM harmonisée (Cackowski et al. 2023; Dewey et al. 2019). Une variante repose sur un seul modèle entraîné avec les coupes dans toutes les orientations ; le volume final est ensuite obtenu de la même manière que dans la première technique (Zuo et al. 2021b). De manière similaire, Liu et al. (2023) ont entraîné leur modèle d'harmonisation en rassemblant les coupes des trois orientations. En outre, en vue de capturer plus d'information sur le volume, le modèle traite trois coupes consécutives. A l'inférence, un volume est généré à partir de chaque orientation en faisant une moyenne pour les coupes appartenant à plusieurs volumes 3D partiels. Le volume final est ensuite obtenu en faisant une moyenne par voxel des trois volumes. De leur côté, Zuo et al. (2021b) ont proposé un apprentissage en deux étapes. Dans une première étape, un modèle 2D est entraîné sur des coupes prises dans toutes les orientations. Ensuite, pour chaque orientation et chaque image d'entraînement, un volume est généré avec le modèle entraîné. Dans une deuxième étape, un modèle de fusion 3D supervisé est alors entraîné avec une fonction de coût par voxel à reproduire les images d'origine à partir des trois volumes synthétisés. Les auteurs pointent des améliorations grâce à cette étape de synthèse 3D mais mentionnent tout de même que leurs résultats pourraient ne pas être optimaux étant donné que le modèle n'est pas entraîné de bout en bout.

Afin de pouvoir exploiter l'information volumique tout en évitant une surcharge de la mémoire GPU, certains travaux reposent sur des modèles 3D entraînés sur des *patches* 3D extraits des images IRM (Chen et al. 2021; Palladino et al. 2020). Zhong et al. (2020) ont proposé un modèle 3D traitant la totalité de cartes d'AF et de DM. Ils ont cependant mis en avant des résultats d'harmonisation similaires avec une architecture 2D, qui a de plus l'avantage de la rapidité de calcul.

2.6. Récapitulatif des types de méthodes

La figure 16 illustre les différentes familles de méthodes que nous avons vues dans ce chapitre. Ces dernières années, beaucoup de travaux de recherches ont porté sur des modèles génératifs non-supervisés. Cela s'explique par le potentiel de ces approches qui permettent d'exploiter les nombreuses bases de données qui sont désormais facilement accessibles dans le monde sans nécessiter de conditions très contraignantes comme des sujets voyageurs sur les différents sites de l'étude. Elles visent de plus à générer des images IRM pouvant être exploitées dans diverses tâches subséquentes comme l'analyse radiologique, l'extraction automatisée de caractéristiques et la prédiction d'informations biologiques. Les développements méthodologiques durant cette thèse ont également porté sur ce type de méthodes (sections 4 et 5).

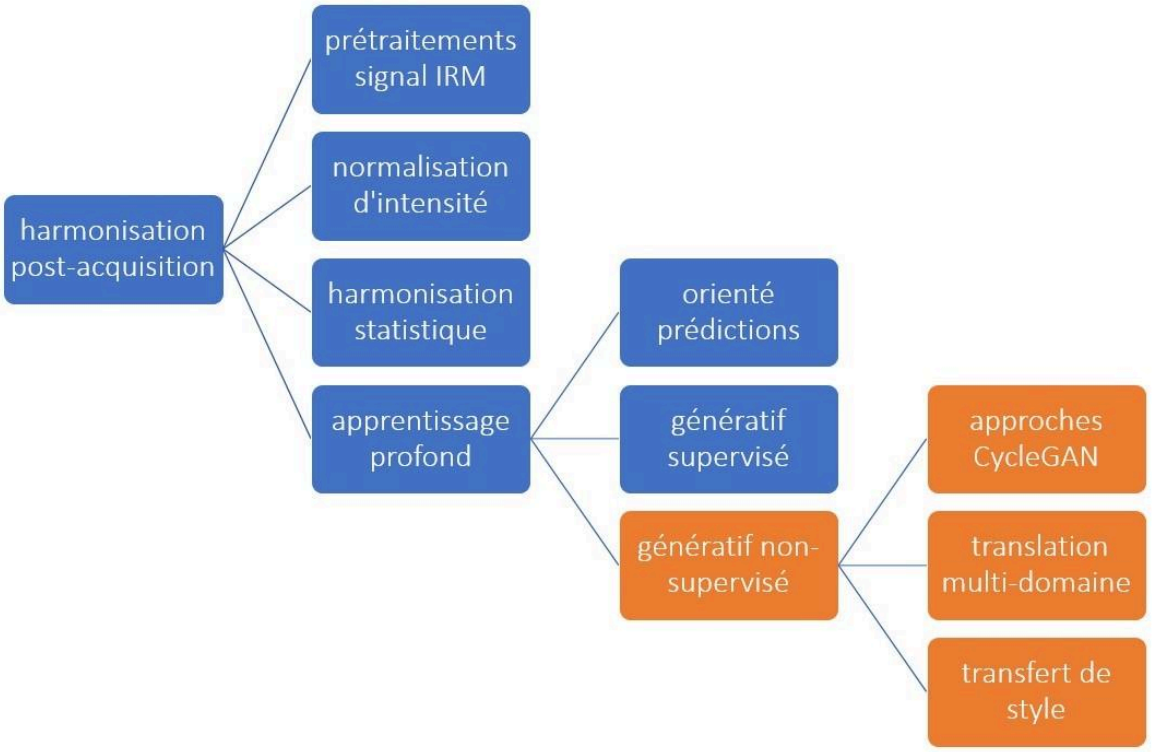


Figure 16 : Diagramme des familles de méthodes d'harmonisation post-acquisition en IRM. Les modèles génératifs non-supervisés, représentés en orange, constituent le principal cadre de recherche de cette thèse.

3. Méthodes d'évaluation des techniques d'harmonisation en IRM : revue et expérimentations

3.1. Introduction

L'objectif général de l'harmonisation peut être exprimé par deux sous-objectifs : éliminer les différences inter-domaines et préserver l'information biologique. Bien que ces objectifs soient simples à appréhender, ils sont abstraits et doivent être précisés pour la mise en place de méthodes d'évaluation. Shinohara et al. (2014) ont proposé des objectifs plus spécifiques pour une approche de normalisation d'intensité, parmi lesquels on trouve les suivants :

1. La normalisation est réplicable.
2. Le rang des intensités est préservé.
3. Les distributions d'intensité pour un tissu d'intérêt sont similaires entre différentes images IRM, pour un même sujet ou pour différents sujets.
4. La normalisation n'est pas influencé par des anormalités biologiques ou de l'hétérogénéité de population.
5. La normalisation n'est pas sensible aux bruits et aux artefacts.
6. La normalisation ne résulte pas en une perte d'information associée à une pathologie ou à d'autres phénomènes biologiques.

Le point 1 est particulièrement important pour les approches complexes contenant des processus stochastiques comme c'est souvent le cas pour l'apprentissage profond. Bien que les auteurs ne pensaient pas spécifiquement à ce type d'approches, les réseaux de neurones sont en effet presque toujours initialisés par échantillons de variables aléatoires. Même au cours de l'apprentissage, certaines techniques comme le *Dropout* (Srivastava et al. 2014) introduisent de l'aléatoire pour éviter le sur-apprentissage. La répétition de plusieurs apprentissages suivi de l'évaluation de la variabilité des résultats d'harmonisation est un moyen d'évaluer la réplicabilité dans ce cas. Le point 2 est une concrétisation de la préservation de l'information individuelle. Si le rang des intensités est préservé, on peut considérer que l'information d'origine est encore présente dans l'image. Le point 3 exprime l'homogénéité souhaitée entre différentes images IRM. Les points 4 et 5 expriment la robustesse nécessaire d'une technique de normalisation, qui doit être capable de traiter des données variables en termes de caractéristiques biologiques mais également en termes de caractéristiques techniques liées à l'acquisition. Le point 7 souligne l'importance de la préservation d'informations d'intérêt, typiquement celles qui sont liées à une pathologie.

Shinohara et al. (2014) se sont toutefois limités à l'analyse de distribution d'intensité pour évaluer des techniques de normalisation. Dans une première partie de ce chapitre, diverses méthodes d'évaluation employées ces dernières années pour l'harmonisation sont expliquées. Dans une deuxième partie, les limites de certaines méthodes et les points d'attention nécessaires à une utilisation efficace sont mis en avant à travers quelques expériences menées dans le cadre de cette thèse.

3.2. Revue des méthodes d'évaluation

3.2.1. Evaluation qualitative

Une manière simple de rendre compte rapidement de la qualité d'une harmonisation pour les modèles génératifs est la visualisation des images avant et après harmonisation à côté d'images du domaine de référence. Cela est particulièrement pratique pour les approches de transfert de style qui reposent sur l'extraction de style associé à une seule image (Figure 17).

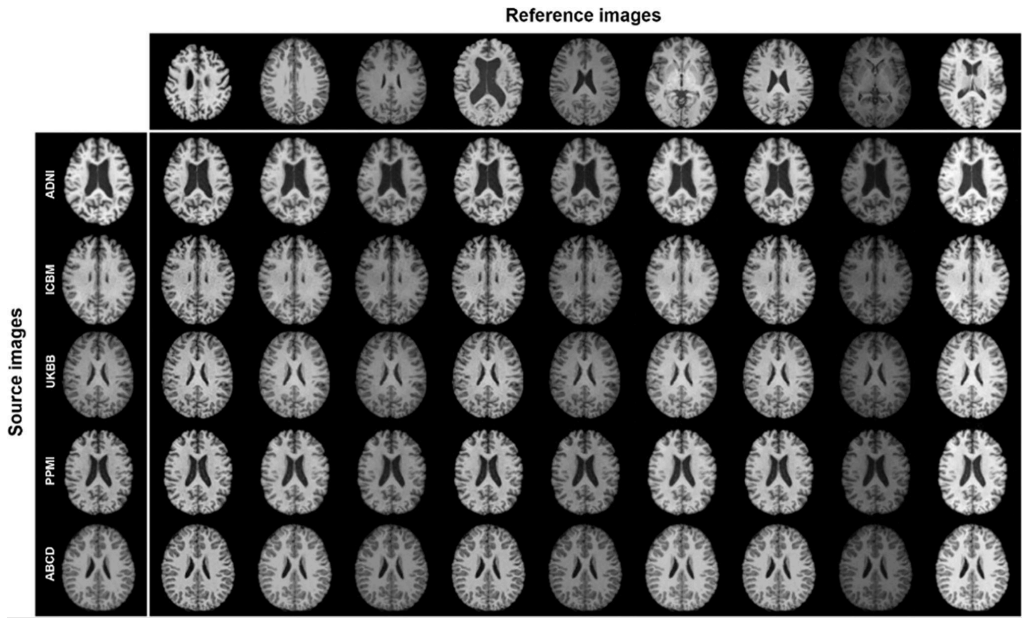


Figure 17 : Coupes avant et après harmonisation par une méthode de transfert de style.
Figure de Liu et al. (2023).

Certaines inspections visuelles portent sur des résultats de segmentation et visent à rendre compte de la consistance inter-sites pour un même sujet (Figure 18) ou de l'évolution des erreurs suite à l'harmonisation (Chen et al. 2021). Des estimations visuelles plus poussées sont parfois mises en place avec des experts radiologues afin de rendre compte du réalisme des images générées (Armanious et al. 2020; Welander et al. 2018).

Pour se rendre compte de la réduction de variabilités entre différents domaines, la visualisation d'images est cependant limitée à un petit nombre d'images et l'analyse ne peut donc porter que sur des différences assez grossières (Figure 17). La représentation par histogrammes d'intensité est un moyen efficace pour représenter des données IRM de différentes images dans un même graphique et d'ainsi repérer d'éventuelles hétérogénéités (Figure 19).

L'utilisation de méthodes de réduction de dimension comme l'ACP ou tSNE (Maaten et Hinton 2008) permet également de regrouper des données IRM dans des graphiques. Une visualisation 2D grâce à des nuages de points permet par exemple d'identifier d'éventuels clusters spécifiques à des sites (Figure 20).

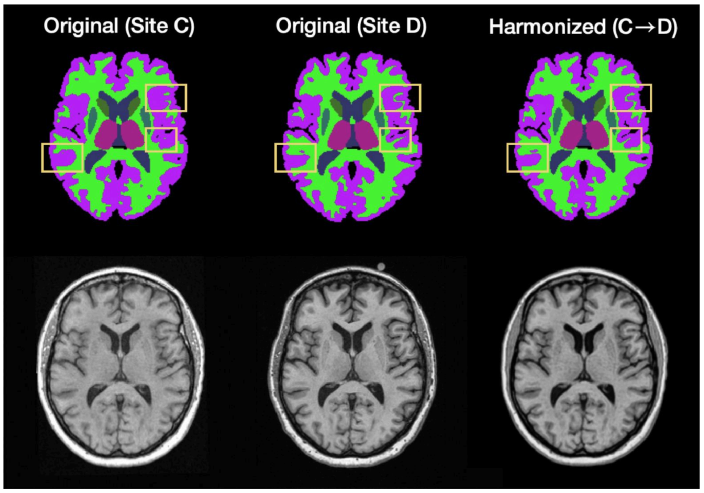


Figure 18 : Comparaison de segmentation sur un sujet voyageur avant et après harmonisation avec CALAMITI. La ligne du bas et la ligne du haut montrent des images T1w et les segmentations SLANT (Huo et al. 2019) correspondantes. Les boîtes jaunes mettent en avant une meilleure consistance des segmentations après harmonisation. Figure de Zuo et al. (2021b).

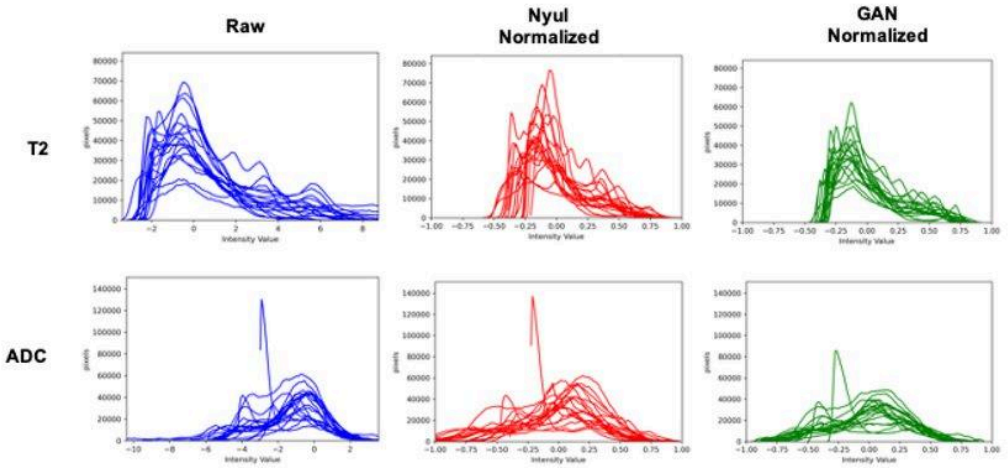
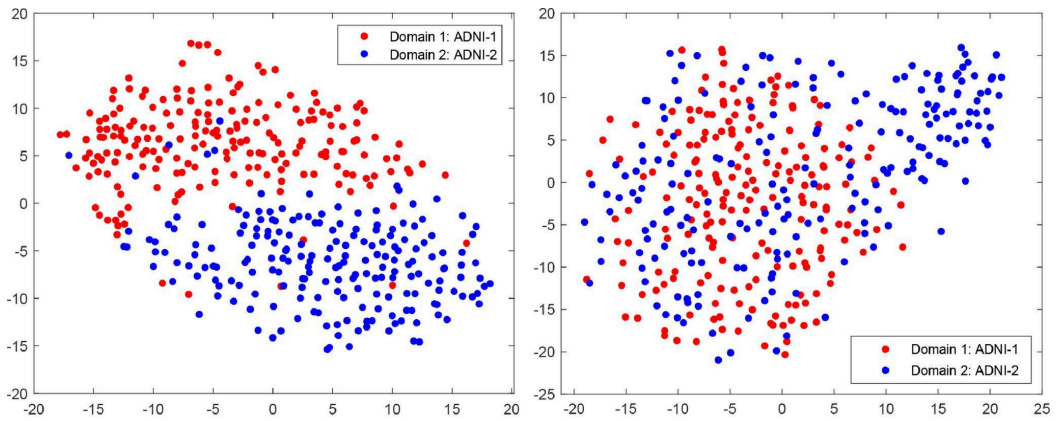


Figure 19 : Histogrammes d'intensité avant et après deux normalisations. Dans chaque sous-figure, chaque ligne correspond à une IRM. Les sous-figures du haut et du bas ont été obtenues avec des images T2w et des coefficients de diffusion apparents, respectivement. Figure de DeSilvio et al. (2021).

3.2.2. Évaluation avec vérité terrain

La plupart des approches d'harmonisation ne requièrent pas de phase d'apprentissage supervisé (section 2). Cela s'explique par la rareté des bases de données avec "vérité terrain" fournissant pour chaque exemple d'entraînement l'harmonisation à obtenir. Il existe néanmoins quelques jeux de données IRM contenant de telles paires. Ils présentent typiquement un petit nombre de sujets pour lesquels une image IRM a été obtenue sur différentes machines. Si le délai entre les différentes acquisitions est suffisamment court (e.g. < 2 mois), on peut utiliser certaines des images comme des vérités terrain. Nous avons déjà mentionné certaines méthodes d'apprentissage automatique qui utilisent ce type de données durant l'apprentissage (section 2.5.2). Des bases de données de sujets voyageurs

ont par ailleurs été utilisées dans de nombreux travaux d'harmonisation afin de disposer de vérités terrain pour l'évaluation, y compris pour des approches non-supervisées.



(a) Before adaptation (b) After adaptation via AD^{2}A

Figure 20 : Visualisation de données d'épaisseurs corticales extraites automatiquement d'images IRM et traitées avec tSNE. Figure de Guan et al. (2021).

3.2.2.1. Évaluation supervisée au niveau voxels

Même en admettant ces vérités terrain, l'évaluation des performances de prédiction de données de grande dimension comme des images IRM n'est pas triviale. Pour estimer les qualités de *reconstruction* d'images, des métriques appliquées voxel-à-voxel comme l'erreur quadratique moyenne ou le *Peak Signal to Noise Ratio* sont couramment employées. Wang et al. (2004) ont pointé les limites de ces métriques qui diffèrent beaucoup du jugement humain dans l'évaluation des erreurs de reconstruction d'images *naturelles*. Les auteurs ont alors proposé le *structural similarity index measure* (SSIM). Cette métrique combine trois termes quantifiant les similarités de luminosité, de contraste et de structure. Elle est appliquée successivement sur des sous-régions de l'image afin de mieux saisir les informations au niveau local et d'être plus cohérente avec la perception humaine.

Ces différentes mesures, développées initialement sur des images naturelles, ont souvent été employées pour évaluer l'harmonisation inter-sites en IRM sur des jeux de données de sujets voyageurs (Cackowski et al. 2023; Dewey et al. 2019; Gao et al. 2019; Liu et al. 2023; Liu et Yap 2021; Qu et al. 2020; Torbati et al. 2021; Zuo et al. 2021b, 2022). En l'absence de sujets voyageurs, Zhong et al. (2020) ont utilisé un appariement basé sur l'âge pour évaluer un modèle d'harmonisation de métriques de diffusion.

Des limites relatives à ce type de métriques ont été mises en avant dans le cas général (Dosselmann et Yang 2011; Palubinskas 2014). Pambrun et Noumeir (2015) ont de plus montré que certaines présomptions faites sur des images *naturelles* pouvaient poser problème dans l'application du SSIM en imagerie médicale. De leur côté, Gourdeau et al. (2022) ont mis en place des expériences montrant des limites potentielles dans l'utilisation du SSIM pour évaluer et comparer des modèles de synthèse d'images médicales. Premièrement, ils insistent sur le biais induit par l'application du SSIM sur des images contenant des valeurs négatives ; cela est fréquent dans le contexte de synthèse d'images médicales où des étapes de normalisation impliquent souvent des valeurs négatives (e.g. Z-score et WhiteStripe, voir sections 2.3.2 et 2.3.3). Deuxièmement, ils mentionnent que la comparaison de méthodes de synthèse d'images médicales est d'autant plus difficile avec

les différentes stratégies de recadrage/remplissage mises en place dans les différentes études. Troisièmement, ils mettent en avant l'influence du paramètre de SSIM fixant la *gamme dynamique* des valeurs de voxel. Enfin, bien que la grande majorité des modèles reposent sur des architectures 2D, ils recommandent d'appliquer le SSIM en 3D quand les images synthétisées sont des volumes. Ravano et al. (2022) ont également montré des limites du SSIM - et d'autres mesures de similarité d'image -, dans le contexte particulier de l'harmonisation inter-sites. Dans leurs expériences, malgré des améliorations des métriques de similarité grâce à un modèle d'harmonisation, des différences inter-sites de volumes cérébraux obtenus avec un logiciel de segmentation sont maintenues. Ils en déduisent donc que ces mesures ne sont pas efficaces pour évaluer la robustesse d'un modèle à des post-traitements et concluent qu'elles peuvent pauvrement refléter l'évolution de la consistance entre domaines.

3.2.2.2. Évaluation supervisée sur des caractéristiques d'image

Pour éviter les écueils mentionnés dans le paragraphe précédent, il est possible de se focaliser uniquement sur des similarités de certaines caractéristiques extraites des images IRM. Des similarités et des différences entre histogrammes ont par exemple été quantifiées avec une corrélation d'histogrammes (Gao et al. 2019) ou une divergence de Jensen-Shannon (Liu et al. 2023), respectivement. Il est à noter que Liu et al. n'ont pas utilisé de sujets voyageurs pour cette expérience mais des sujets appariés en fonction de leur âge et de leur sexe.

D'autres études contiennent des analyses de différences de segmentation avec des sujets voyageurs pour rendre compte du potentiel intérêt de l'harmonisation. Des chercheurs ont par exemple calculé des indices de Dice entre des segmentations FSL-FAST (Zhang et al. 2001) d'images IRM de différents sites pour un même sujet (Liu et Yap 2021; Torbati et al. 2021). Zuo et al. (2021b) et Dewey et al. (2019) ont quantifié de la même manière des similarités de segmentations, sur 9 régions obtenues avec SLANT (Huo et al. 2019) et sur 5 structures cérébrales extraites avec une autre pipeline automatisée, respectivement. De leur côté, Gebre et al. (2023) ont utilisé des corrélations intra-classe (Fisher 1992) sur 34 mesures FreeSurfer d'épaisseurs corticales pour évaluer la consistance des segmentations de sujets voyageurs (Figure 21).

3.2.3. Étude de différences de distributions entre domaines

L'évaluation des méthodes d'harmonisation ne peut pas se limiter à des jeux de données de sujets voyageurs. Nous avons vu précédemment que des "pseudo-appariements" ont été mis en place pour contourner cette limitation (Liu et al. 2023; Zhong et al. 2020). Néanmoins, une pratique plus courante consiste en l'analyse de différences entre des distributions de données correspondant à différents domaines (différents sites typiquement). Fortin et al. (2017) ont par exemple appliqué des tests de Student à deux échantillons pour identifier les voxels significativement différents entre les domaines sur des cartes d'AF et de DM. Avec le même objectif, Tian et al. (2022) ont utilisé des ANOVA au niveau des voxels sur des cartes de volumes de MG.

Ces tests montrent en général que moins de différences sont significatives après harmonisation. Ils sont néanmoins plus souvent employés avec des caractéristiques extraites des images, y compris pour l'évaluation de modèles harmonisant au niveau des voxels - ce qui implique alors une phase d'extraction de caractéristiques postérieure à l'harmonisation. L'étude de mesures volumétriques corticales et sous-corticales est un

exemple assez courant, que ce soit avec des tests à deux échantillons comme celui de Student ou de Kolmogorov-Smirnov (KS) (Garcia-Dias et al. 2020; Gebre et al. 2023), avec des ANOVA et des tests de Bartlett sur de multiples domaines (Fortin et al. 2018) ou des scores de Cohen pour mesurer la taille des effets (Liu et al. 2023). Chen et al. (2022) ont également étudié l'harmonisation d'épaisseurs corticales, mais ont ajouté des simulations d'effets de site relatives aux moyennes, variances et covariances des mesures. Les auteurs se sont en particulier intéressés aux effets de sites sur les covariances (Figure 22). Des différences portant sur des radiomiques ont aussi été étudiées avec des tests U de Mann-Whitney (Chang et al. 2022) ou des tests t de Welch et des tests de KS (Fatania et al. 2022). Torbati et al. (2021) ont de manière similaire étudié des différences inter-sites sur des segmentations du cerveau en MG, MB et LCS. Pomponio et al. (2020) ont de leur côté utilisé des diagrammes en boîte pour rendre compte de différences inter-sites de volumes hippocampiques.

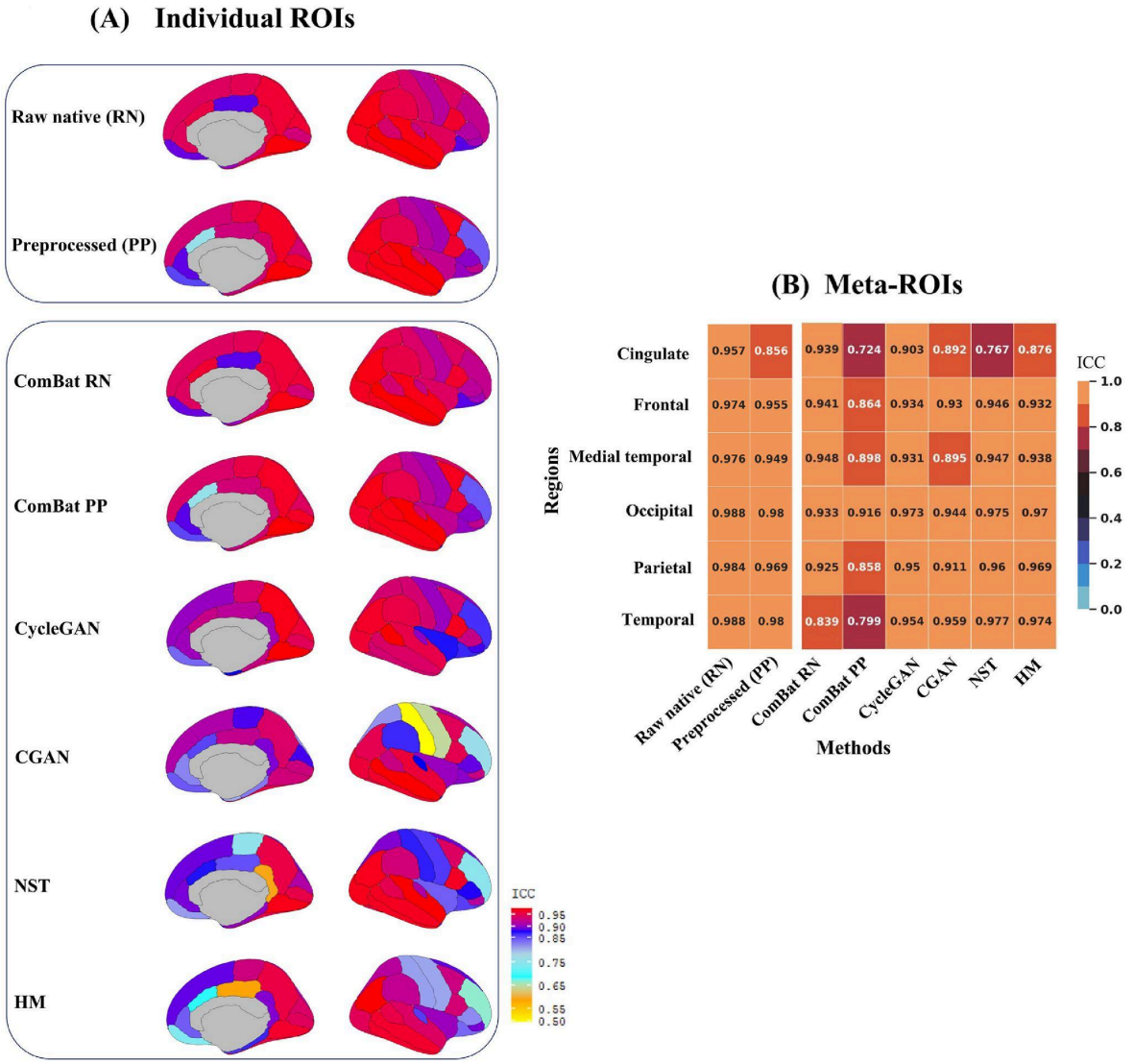


Figure 21 : Cartes de chaleur de corrélations intra-classes (ICC) sur des sujets voyageurs. Les résultats sont montrés pour les images natives (Raw native), les images prétraitées (Preprocessed) et les images harmonisées avec différentes méthodes. **(A)** Régions d'intérêt (ROIs) individuelles segmentées par FreeSurfer. **(B)** ROIs regroupées en six meta-ROIs. Figure de Gebre et al. (2023).

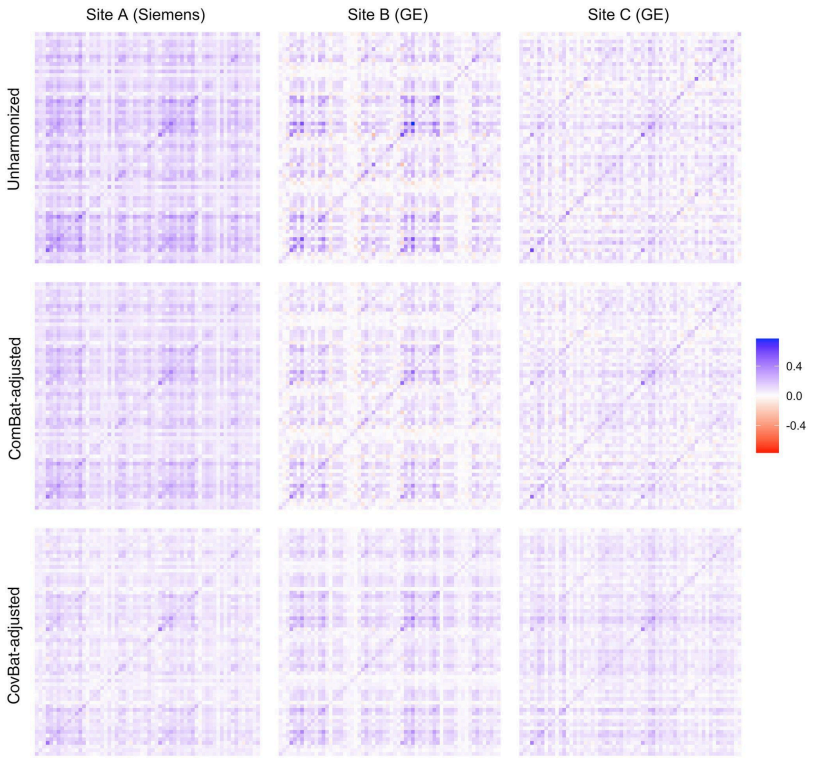


Figure 22 : Matrices de covariance de mesures d'épaisseurs corticales. Les covariances ont été obtenues après que les mesures ont été *résidualisées* sur l'âge, le sexe et le diagnostique clinique. Les colonnes et les lignes correspondent respectivement au site d'acquisition et à une méthode d'harmonisation (*Unharmonized* correspond aux données brutes). Figure de Chen et al. (2022).

Avec la même idée, la prédiction du domaine d'origine des images IRM est une approche assez courante (Cackowski et al. 2023; Chen et al. 2022; Nguyen et al. 2018; Wachinger et al. 2021). L'idée est qu'un modèle doit être moins performant dans la distinction des domaines si l'harmonisation a été efficace.

L'hypothèse sous-jacente de ces évaluations est que les différences dans des distributions de caractéristiques entre domaines sont des différences techniques, non pertinentes, que l'harmonisation doit supprimer. Un emploi pertinent de ces évaluations implique donc que les populations étudiées soient semblables biologiquement. Il ne serait par exemple pas attendu qu'une population de personnes âgées présentent des distributions d'épaisseurs corticales similaires à celles d'une population d'enfants. Si les groupes comparés sont hétérogènes, une correction de certaines covariables peut être appliquée préalablement (Chen et al. 2022; Pomponio et al. 2020). Même si l'on arrive à éviter les biais de populations, ces approches présentent cependant l'inconvénient de ne rendre compte que de différences inter-domaines et non de conservations d'informations individuelles.

3.2.4. Étude de motifs biologiques

Pour évaluer la qualité d'une harmonisation en l'absence de sujets voyageurs, il est donc nécessaire de prendre en compte l'évolution des motifs biologiques dans les données. Fortin et al. (2017) ont par exemple mis en place une expérience pour évaluer la réplicabilité des voxels d'AF associés à l'âge des sujets suite à l'harmonisation. Pour identifier ces voxels, un

test statistique inférentiel à partir d'une régression linéaire est utilisé. D'autres études se sont penchés sur des motifs de vieillissement, notamment en étudiant des corrélations linéaires (coefficient de Pearson) entre l'âge et une mesure globale d'épaisseur corticale (Fortin et al. 2018) ou entre l'âge et des mesures d'AF le long de faisceaux de MB (Zhong et al. 2020). Zhong et al. (2020) se sont aussi intéressés aux différences inter-genre sur l'AF avec des scores de Cohen. De leur côté, Dewey et al. (2019) ont utilisé des données longitudinales pour étudier les différents motifs d'atrophie de la MG corticale entre les protocoles (Figure 23).

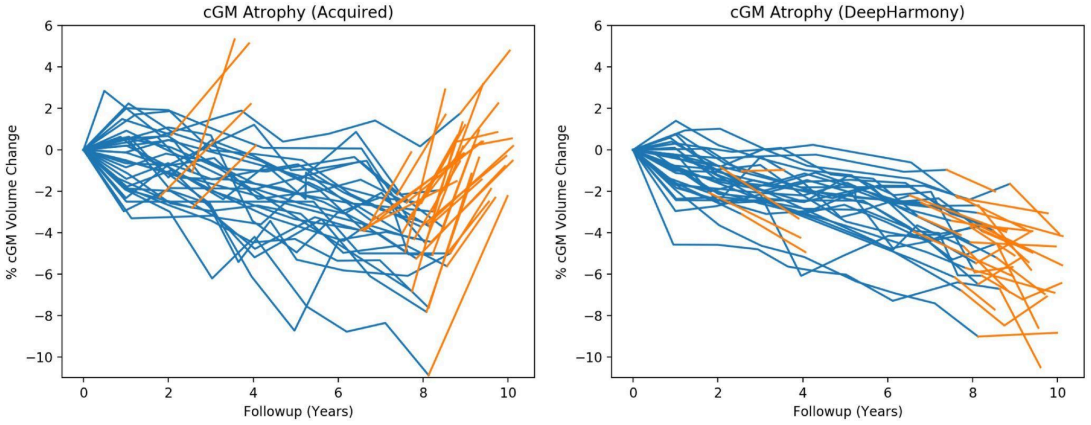


Figure 23 : Évolution de volumes de matière grise corticale. Les volumes sont relatifs au volume intracrânien de la première visite et sont exprimés par rapport au volume relatif de matière grise corticale de la première visite. Chaque trajectoire correspond à un sujet et la couleur indique le protocole d'acquisition. L'image de gauche et l'image de droite correspondent respectivement aux résultats avant et après harmonisation. Figure de Dewey et al. (2019).

La conservation de biomarqueurs pathologiques est également un enjeu majeur de l'harmonisation. Bayer et al. (2022) ont ainsi déterminé un critère de typicité d'épaisseur corticale par région d'intérêt - utilisant le nombre d'écart-types par rapport à la moyenne - pour déterminer pour chaque image des régions potentiellement atypiques. Ils ont ensuite comparé les résultats d'application de ce critère entre des sujets autistes et des sujets contrôles. Liu et al. (2023) se sont eux intéressés à la maladie d'Alzheimer avec des scores de Cohen pour comparer les volumes hippocampiques de sujets malades et de sujets sains.

3.2.5. Performances de modèles de prédiction

Nous avons précédemment évoqué l'utilisation de méthodes d'apprentissage automatique supervisé sur des données IRM (section 1.3). Des études ont mis en avant que ce type d'approches pouvait souffrir de biais avec des jeux de données multicentriques. Typiquement, un modèle prédictif peut être performant sur des données issues du domaine d'entraînement mais significativement moins bon pour généraliser à d'autres données (Zech et al. 2018). Ces approches sont alors intéressantes pour évaluer la capacité des méthodes d'harmonisation à conserver des informations d'intérêt, voire à améliorer les prédictions grâce à une réduction de la variabilité inter-sites.

La prédiction de l'âge de sujets sains en est un bon exemple puisqu'elle a servi dans de nombreuses études pour rendre compte de l'apport de l'harmonisation. Ainsi, Fortin et al. (2018) ont mis en place des régressions linéaires et des machines à vecteur de supports (SVM) non-linéaires pour prédire l'âge à partir de mesures d'épaisseurs corticales et ont

reporté les performances avec différentes méthodes d'harmonisation. Pomponio et al. (2020) ont fait de même avec un réseau de neurones à couches denses appliqué à des volumes cérébraux. Des études proposant des harmonisations au niveau voxel ont également utilisé des architectures profondes pour prédire l'âge (Bashyam et al. 2022; Liu et al. 2023; Robinson et al. 2020). Cette validation a aussi été employée pour l'évaluation du modèle d'harmonisation orienté prédiction de Dinsdale et al. (2021).

Tout comme l'âge, le sexe est une information qui est presque systématiquement fournie dans les bases de données IRM. Sa prédiction avec de l'apprentissage automatique a donc également été mise en place pour valider des méthodes d'harmonisation : sur des mesures d'épaisseurs corticales avec une forêt d'arbres décisionnels (Chen et al. 2022) ; sur l'image IRM avec une SVM non-linéaire appliquée à des composantes principales d'une ACP (Nguyen et al. 2018) ou avec un CNN (Robinson et al. 2020).

Comme évoqué précédemment, la conservation d'informations pathologiques est primordiale. Différents classifieurs binaires de diagnostics cliniques ont alors été employés dans des études d'harmonisation :

- alzheimer :
    - seuillage de l'intensité moyenne des voxels hippocampiques (Fortin et al. 2016)
    - forêt d'arbres décisionnels appliqué à des mesures d'épaisseurs corticales (Chen et al. 2022)
    - modèle d'harmonisation orienté prédiction au niveau voxel (Wang et al. 2022)
- schizophrénie :
    - réduction de dimensions d'une ACP sur l'image IRM et SVM non-linéaire (Nguyen et al. 2018)
    - modèle d'harmonisation orienté prédiction au niveau voxel (Wang et al. 2022)
- autisme : SVM non-linéaire appliqué à des radiomiques (Cackowski et al. 2023)
- tumeurs : sélection de caractéristiques sur des radiomiques et SVM pour le stade du gliome (Gao et al. 2019) et le cancer du col de l'utérus (Chang et al. 2022)

L'évaluation d'harmonisation basée sur des performances d'un modèle de segmentation basé sur de l'apprentissage automatique est plus rare car l'acquisition de labels fiables requiert une annotation manuelle d'un expert, souvent coûteuse en temps. Certains chercheurs ont tout de même suivi cette approche afin de rendre compte de l'évolution de la précision dans la localisation d'hyper-intensités de MB (Palladino et al. 2020) ou de cancer de la prostate (DeSilvio et al. 2021). Chen et al. (2021) ont de leur côté utilisé les données du challenge NeoBrainS12 (Išgum et al. 2015) pour évaluer la capacité de leur modèle d'harmonisation à améliorer la segmentation d'images IRM cérébrales de nouveaux-nés en 8 classes (Figure 24). Pour une seconde expérience de segmentation, les auteurs disposaient d'images IRM de sujets âgés de 24 mois et d'autres de sujets âgés de 6 mois. En partant de l'observation que les premières présentaient de meilleurs contrastes de tissus, ils ont postulé qu'un logiciel de segmentation existant comme FreeSurfer pouvait être utilisé sur celles-ci mais pas sur les secondes. Ils ont alors entraîné un modèle à reproduire les segmentations de MG, MB et LCS de FreeSurfer sur les images des sujets âgés de 24 mois et l'ont testé sur celles de ceux âgés de 6 mois avant et après harmonisation vers les premières. Des segmentations manuelles ont ensuite permis de rendre compte de l'apport de l'harmonisation dans cette expérience. Dinsdale et al. (2021) ont également utilisé un logiciel de segmentation automatique de MG, MB et LCS (FSL-FAST (Zhang et al. 2001)) pour générer des pseudo-vérités terrain et ainsi évalué l'apport de leur modèle d'harmonisation orienté prédiction.

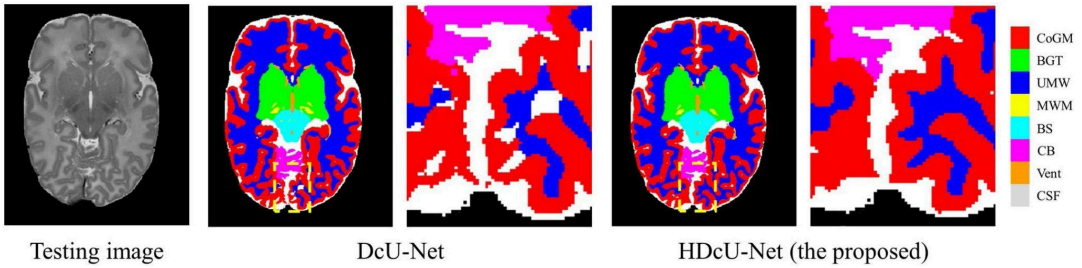


Figure 24 : Évolution de segmentations du cerveau après une harmonisation CycleGAN. Le cerveau est segmenté en 8 classes avec un réseau de neurones type U-net (DcU-Net). La segmentation après harmonisation vers le domaine des images IRM d'entraînement est également montrée (HDcU-Net). Des vues zoomées pointent certaines différences notables. Figure de Chen et al. (2021).

3.3. Limites de méthodes d'évaluation dans la littérature

Comme vu dans la section 3.2.2.1, l'utilisation de métriques de similarité d'images sur des sujets voyageurs est la méthode d'évaluation la plus courante des approches d'harmonisation dans la littérature. Dans cette partie, certaines limites de ces approches sont illustrées à travers quelques expériences menées sur le jeu de données de sujets voyageurs de la base SRPBS (SRPBS_TS, Tanaka et al. 2021). Cette base contient des images IRM T1w de cerveaux d'hommes sains âgés entre 24 et 32 ans. Nous avons inclus onze machines sur lesquelles chaque sujet avait été scanné. Nous avons supprimé deux images à cause d'un problème de correspondance entre les données et les métadonnées (images des sites *YC2* et *UTO*, respectivement pour les sujets 2 et 3).

3.3.1. Expérience 1 : Impact de l'extraction de la boîte crânienne sur les métriques de similarité

3.3.1.1. Introduction

Comme vu dans la section 2.2.3, l'extraction de la boîte crânienne est courante dans les pipelines d'harmonisation mais un certain nombre de chercheurs ont fait le choix de la conserver. Cackowski et al. (2023) ont comparé leur approche d'harmonisation d'images IRM cérébrales - qui inclut une extraction de la boîte crânienne - à celle de Zuo et al. (2021a) - qui n'en inclut pas. Pour cela, ils ont utilisé des sujets voyageurs de la base OASIS-3 (LaMontagne et al. 2019) et ont calculé les SSIMs entre les images d'un même sujet mais de différents scanners après harmonisation avec leur méthode. En comparant avec les résultats rapportés dans l'article de Zuo et al., Cackowski et al. ont mis en avant de meilleurs SSIMs avec leur méthode. Un des potentiels biais de cette comparaison est l'extraction de la boîte crânienne qui n'est réalisée qu'avec une des deux méthodes. Dans cette expérience, nous avons étudié l'impact de l'extraction de la boîte crânienne sur le SSIM.

3.3.1.2. Méthodes

Pour cette expérience, nous avons appliqué les prétraitements suivants aux images de SRPBS_TS :

1. Extraction de la boîte crânienne avec HD-BET (Isensee et al. 2019).

2. Recalage linéaire (douze degrés de liberté) vers MNI152 1mm^3 avec FSL-FLIRT (Jenkinson et al. 2002).
3. Application des transformations de l'étape 2 aux images brutes (i.e. avec la boîte crânienne).

Nous avons ensuite calculé le SSIM (Wang et al. 2004) pour chaque sujet entre chaque paire de site en fixant la gamme dynamique à un million. Nous avons effectué ces calculs sur les images IRM issues de l'étape 2 (sans boîte crânienne) et celles issues de l'étape 3 (avec boîte crânienne).

3.3.1.3. Résultats et discussion

Les résultats de l'expérience sont présentés dans le Tableau 1. Sans la boîte crânienne, une nette augmentation du SSIM est obtenue (+ 0.153). Cette augmentation est importante par rapport aux différences communément mises en avant dans la littérature.

Tableau 1 : SSIMs intra-sujets sur la base SRPBS_TS avec et sans la boîte crânienne.

	avec boîte crânienne	**sans boîte crânienne**
SSIM [1]	0.684 ± 0.322	0.837 ± 0.169

[1] Le SSIM est exprimé comme moyenne ± écart-type.

Cet exemple de biais rendant difficile la comparaison de méthodes de synthèse d'images est plus généralement lié à la gestion de l'arrière-plan - qui peut être plus ou moins grand suivant les stratégies de rognage/remplissage et/ou qui peut prendre différentes valeurs - qui a un impact non négligeable sur le SSIM (Gourdeau et al. 2022).

3.3.2. Expérience 2 : Impact du paramètre de gamme dynamique

3.3.2.1. Introduction

Très peu d'études d'imagerie médicale évoquent le paramètre définissant la gamme dynamique des valeurs de pixel/voxel pour le SSIM. Pourtant, les gammes de valeurs sont rarement standardisées et il est donc difficile de fixer une valeur pertinente par défaut pour ce paramètre. Liu et al. (2023) ont proposé une méthode d'harmonisation d'images cérébrales T1w. Une de leurs validations consiste en l'application du SSIM à la cohorte de sujets voyageurs de Tong et al. (2020). Ils y comparent les SSIMs obtenus avant et après que les images ont été harmonisées avec leur modèle. Par ailleurs, aucune normalisation d'intensités - excepté une correction d'inhomogénéités - n'est évoquée dans leurs prétraitements, ni aucune information sur la gamme dynamique du SSIM. Or, l'harmonisation peut potentiellement modifier significativement la gamme des valeurs de voxel et utiliser une même gamme dynamique peut alors biaiser les résultats. Dans cette expérience, nous avons illustré ce problème avec une simple modification apportée aux images et proposé un petit ajustement visant à contourner le biais.

3.3.2.2. Méthodes

Nous avons réutilisé les images de l'expérience précédente avec boîte crânienne (section 3.3.1.2) et leur avons appliqué un traitement consistant simplement en une division par 2

des intensités de voxel. Nous avons ensuite calculé les SSIMs pour chaque sujet avant et après ce traitement, en gardant la gamme dynamique fixée à un million.

Comme notre hypothèse était que la réduction de l'échelle d'intensités allait permettre une augmentation du SSIM, nous avons également testé d'ajuster la gamme dynamique du SSIM en fonction de chaque paire d'images ; nous avons pour cela utilisé le 99e percentile de l'ensemble des intensités des deux images.

3.3.2.3. Résultats et discussion

Les résultats sont présentés dans le Tableau 2. La réduction de l'échelle d'intensités a effectivement permis une augmentation du SSIM (+ 0.020). Bien qu'elle ne soit pas très importante - notamment par rapport à celle constatée en retirant la boîte crânienne (section 3.3.1.3) -, elle aurait pu l'être avec une réduction plus forte des intensités. Typiquement, une normalisation min-max peut réduire drastiquement une échelle d'intensités d'IRM. Une telle normalisation est d'ailleurs appliquée en prétraitements dans le code Github de la méthode de Liu et al. (2023)[1], bien qu'elle ne soit pas mentionnée dans l'article.

Tableau 2 : SSIMs intra-sujets sur la base SRPBS_TS avant et après réduction des échelles d'intensités.

	images de base	**images avec échelle d'intensités réduite**
SSIM avec gamme dynamique fixe [1]	0.684 ± 0.323	0.704 ± 0.316
SSIM avec gamme dynamique ajustée [1]	0.363 ± 0.194	0.363 ± 0.194

[1] Le SSIM est exprimé comme moyenne ± écart-type.

On remarque également dans le Tableau 2 que les SSIMs sont égaux avec une gamme dynamique ajustée. Cela est souhaitable car il n'est pas attendu qu'une simple diminution de l'échelle d'intensités permette une augmentation des métriques de similarité. Gourdeau et al. (2022) ont proposé un ajustement similaire pour le calcul du SSIM dans le contexte de la synthèse d'images médicales.

3.3.3. Expérience 3 : Adéquation entre le SSIM et la similarité volumétrique

3.3.3.1. Introduction

Ravano et al. (2022) ont expérimenté différentes méthodes d'harmonisation d'images cérébrales T1w. À partir de deux jeux de données de sujets voyageurs, ils ont mis en place deux méthodes d'évaluation : des métriques de similarité d'image - parmi lesquelles un SSIM - et une comparaison de segmentations MorphoBox (Schmitter et al. 2015). Les résultats montrent que, bien que les métriques de similarité aient été améliorées, les segmentations n'étaient pas forcément plus consistantes après harmonisation.

[1] https://github.com/USC-LoBeS/style_transfer_harmonization accédé le 15/07/2023

Dans cette expérience, nous avons évalué l'adéquation entre le SSIM et des similarités de segmentations automatiques de MG, de MB et de LCS.

3.3.3.2. Méthodes

En utilisant les mêmes images IRM que dans l'expérience précédente (section 3.3.2.2), nous avons cette fois appliqué une normalisation min-max, qui est une approche classique dans l'analyse d'images IRM (section 2.3.1).

Nous avons de nouveau calculé les SSIMs intra-sujet, en utilisant l'ajustement de la gamme dynamique (section 3.3.2.2), avant et après la normalisation min-max. Nous avons ensuite calculé pour chaque image IRM les volumes de MG, de MB et de LCS avec des segmentations SPM12.

3.3.3.3. Résultats et discussion

La normalisation min-max a permis une nette augmentation du SSIM intra-sujet moyen (Tableau 3, +0.196). Cela confirme la pertinence de cette approche, notamment pour l'harmonisation. Cependant, la normalisation n'a pas permis une homogénéisation des volumes estimés par SPM12 (Figure 25).

Tableau 3 : SSIMs intra-sujets sur la base SRPBS_TS avant et après normalisation min-max.

	avant min-max	**après min-max**
SSIM [1]	0.363 ± 0.194	0.559 ± 0.119

[1] Le SSIM est exprimé comme moyenne ± écart-type.

Ces résultats vont donc dans le sens des observations de Ravano et al. (2022) qui mettaient en avant une inadéquation entre les évolutions de métriques de similarité d'image et de consistances de segmentation. Ce constat est particulièrement important puisqu'il suggère que ce type de métriques peut être un mauvais indicateur de la capacité des modèles à favoriser la consistance d'applications subséquentes entre les domaines.

3.3.4. Conclusion

Nous avons vu à travers ces trois expériences différentes limites dans l'application du SSIM pour la validation de l'harmonisation sur des sujets voyageurs. Ces limites portent notamment sur les prétraitements puisque l'on voit que l'extraction de la boîte crânienne et un changement dans les gammes d'intensité - qui sont courants en IRM - ont un impact non négligeable sur la métrique. Cela suggère que la comparaison directe de modèles d'harmonisation impliquant différents prétraitements est critiquable.

Un autre élément à considérer pour la comparaison de méthodes est le paramétrage du SSIM ; nous avons en effet observé que la modification de la gamme dynamique avait une influence significative sur les résultats. Pour éviter de potentielles ambiguïtés, les futures études devraient alors expliciter ce paramètre et/ou fournir le code utilisé.

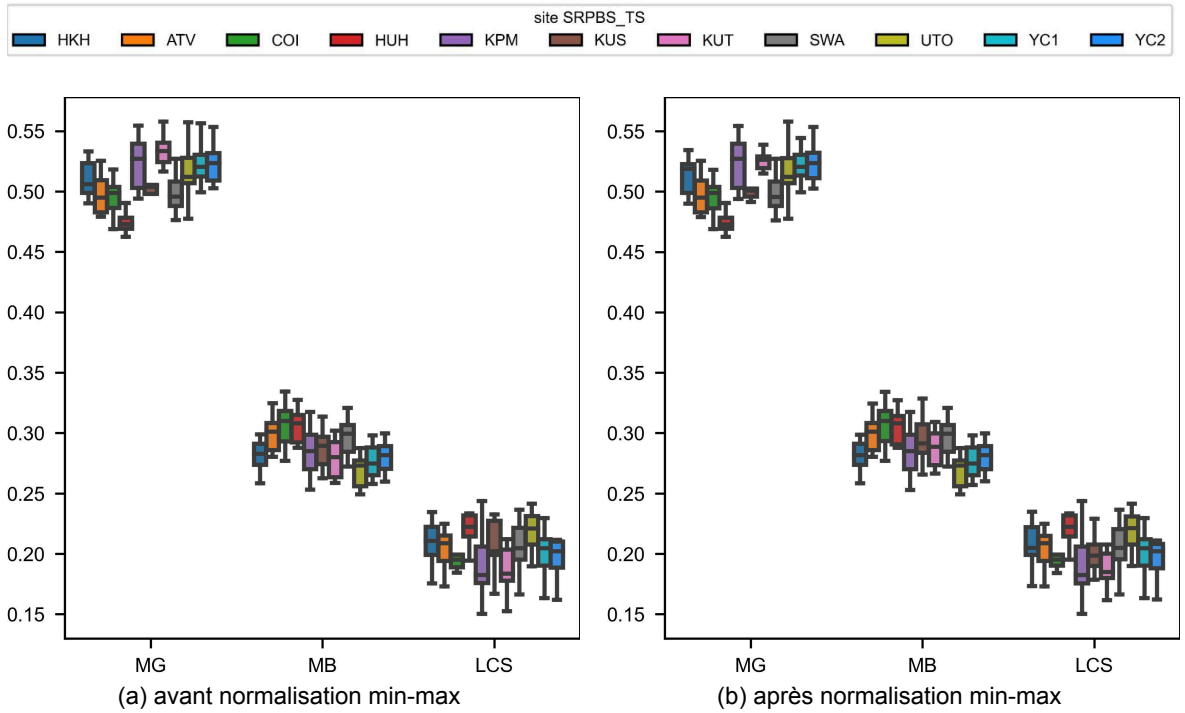


Figure 25 : Estimations de volumes cérébraux sur la base SRPBS_TS avant et après normalisation min-max. Les volumes sont estimés avec SPM12 et divisés par le volume intracrânien total. MG : matière grise ; MB : matière blanche ; LCS : liquide cérébrospinal.

Dans tous les cas, les résultats de l'expérience 3 suggèrent que la valeur du SSIM n'est pas forcément corrélée à la consistance de résultats d'analyses ultérieures. Comme indiqué par Ravano et al. (Ravano et al. 2022), ces évaluations ne sont donc pas suffisantes dans le contexte de l'harmonisation. Il doit être noté que les deux études que nous avons prises en exemple pour des utilisations critiquables de ce type d'évaluations ont également proposé d'autres expériences qui montrent l'intérêt de leurs modèles d'harmonisation (Cackowski et al. 2023; Liu et al. 2023).

4. Un modèle d'apprentissage profond tridimensionnel pour une harmonisation inter-sites d'images IRM structurelles du cerveau : validation approfondie avec un jeu de données multicentrique

Auteurs : Vincent Roca, Grégory Kuchcinski, Jean-Pierre Pruvo, Dorian Manouvriez, Xavier Leclerc et Renaud Lopes

Résumé

Dans les études d'IRM multicentriques, agréger les données d'imagerie peut introduire des variabilités liées au site et peut donc biaiser les analyses ultérieures. Pour harmoniser les distributions d'intensité dans un jeu de données multicentrique, des méthodes d'apprentissage profond non-supervisées peuvent être employées. Ici, nous avons développé un modèle basé sur des réseaux antagonistes génératifs consistants au cycle pour l'harmonisation d'images cérébrales T1w. Contrairement aux précédents travaux, il a été conçu pour traiter des images complètes du cerveau en 3D de manière stable tout en optimisant les ressources de calcul. En utilisant six jeux de données IRM d'adultes sains (n = 1525 au total) avec différents paramètres d'acquisition, nous avons testé le modèle avec (i) trois harmonisations par paire avec des effets de site de taille variable, (ii) une harmonisation globale des six jeux de données avec des différences de distribution d'âge et (iii) un jeu de données de sujets voyageurs. Nos résultats sur des distributions d'intensité, des volumes cérébraux, des IQMs et des caractéristiques radiomiques indiquent que les caractéristiques IRM sur les différents sites ont été homogénéisées efficacement. Ensuite, des expériences de prédiction d'âge et la corrélation observée entre les volumes de MG et les âges montrent que, grâce à une stratégie d'entraînement appropriée et malgré les différences biologiques entre les populations des jeux de données, le modèle a renforcé des motifs biologiques. De plus, des analyses radiologiques sur des images harmonisées attestent de la conservation d'informations radiologiques des images originales. La robustesse du modèle d'harmonisation (telle que jugé par les différents jeux de données et les différentes métriques) démontre son potentiel pour des applications dans des études multicentriques rétrospectives.

Mots clés : IRM cérébrale ; harmonisation ; CycleGAN ; volumétrie cérébrale ; âge cérébral

4.1. Introduction

Nous avons vu dans la section 1.1 que l'IRM cérébrale est communément employée dans l'étude de maladies neurologiques ou psychiatriques. Dans la section 1.2, nous avons expliqué l'intérêt des études multicentriques en IRM et introduit la problématique de la variabilité inter-sites. Nous avons ensuite présenté dans la section 2 différents types de méthode pour l'harmonisation rétrospective.

Dans ce chapitre, nous nous focalisons sur l'utilisation de modèles d'apprentissage profond non-supervisés pour l'harmonisation d'images IRM cérébrales T1w. CycleGAN (Zhu

et al. 2017) est certainement la méthode la plus validée. Bien qu'elle requière un nouvel entraînement pour chaque nouveau site, elle a montré des résultats encourageants dans l'harmonisation de différentes modalités d'images et différentes séquences IRM (Chen et al. 2021; Nguyen et al. 2018; Palladino et al. 2020), particulièrement en comparaison d'approches statistiques (Gebre et al. 2023). D'autres approches d'apprentissage profond ont été proposées pour éviter un réentraînement à chaque nouveau site mais n'utilisent pas l'information du site (Liu et al. 2021), requièrent deux différentes séquences IRM pour chaque sujet (Zuo et al. 2021b) ou sont limitées en termes de validation (Cackowski et al. 2023; Liu et Yap 2021).

De plus, comme détaillé dans la section 2.5.3.4, toutes ces études sont basées sur des modèles d'apprentissage profond 2D (potentiellement répétés sur les trois axes) ou des modèles 3D appliqués à des patches de petit volume. Les auteurs ne discutent pas ou très peu ce choix en général ou le justifient par des contraintes matérielles.

A côté de ces défis méthodologiques, l'évaluation de la qualité des résultats d'harmonisation est également un aspect clé. La section 3.2 propose une large revue des méthodes employées dans la littérature : métriques de similarité basées sur des sujets voyageurs pour lesquels des vérités terrain peuvent être présumées (Cackowski et al. 2023; Gao et al. 2019; Liu et Yap 2021; Nguyen et al. 2018; Zuo et al. 2021b), prédiction du sexe (Nguyen et al. 2018; Robinson et al. 2020), de l'âge (Bashyam et al. 2022; Robinson et al. 2020), de pathologies (Cackowski et al. 2023; Gao et al. 2019; Nguyen et al. 2018) et de segmentations de tissus cérébraux (Chen et al. 2021; Liu et al. 2021; Palladino et al. 2020). Toutefois, dans la plupart des travaux, les validations ne sont pas assez approfondies pour rendre compte de la robustesse des modèles à diverses applications. De plus, des effets de site ont été reportés pour des caractéristiques extraites d'images cérébrales T1w : volumes de tissu (Gunter et al. 2021), caractéristiques radiomiques (Acquitter et al. 2022) et IQMs (Esteban et al. 2017).

Dans ce chapitre, nous proposons un modèle CycleGAN 3D l'harmonisation inter-sites d'images cérébrales T1w qui permet le traitement d'images 3D tout en préservant la stabilité et en optimisant les ressources de calcul. En utilisant six jeux de données, nous avons mesuré des différences inter-sites en distribution d'intensité, volumes cérébraux, caractéristiques radiomiques et IQMs pour évaluer notre approche. Nous avons utilisé l'âge pour quantifier des motifs biologiques avec la prédiction d'âge et la corrélation entre l'âge et le volume de MG, et nous avons utilisé des échelles spécifiques pour coter la préservation de motifs radiologiques. Nous avons testé le modèle sur des cohortes avec différentes distributions d'âge et avons pu les harmoniser efficacement tout en évitant la sur-correction grâce à une stratégie d'entraînement appropriée. De plus, nous avons validé la qualité des reconstructions avec un jeu de données de sujets voyageurs.

4.2. Matériels et méthodes

4.2.1. Jeux de données

4.2.1.1. Jeux de données indépendants

Nous avons obtenu des images cérébrales T1w 3D à partir de trois sources de partage de données : IXI[2], OASIS-3 (LaMontagne et al. 2019), NKI-RS (Nooner et al. 2012) et

[2] https://brain-development.org/ixi-dataset/ accédé le 15/01/2022

NMorphCH et avons ensuite créé six jeux de données d'images acquises avec différentes machines (Tableau 4). Tous les participants étaient présents dans un seul jeu de données. Comme spécifié dans les protocoles des études, tous les participants étaient des contrôles-sains. Chaque participant avait donné son consentement informé au site de l'étude locale et chaque contribution fût approuvée éthiquement.

Tableau 4 : Caractéristique des participants et des scanners des six jeux de données indépendants de la section 4.

Nom du jeu de données	Site1	Site2	Site3	Site4	Site5	Site6
Etude	IXI	IXI	OASIS-3	OASIS-3	NKI-RS	NMorphCH
Images IRM, n	309	176	984	453	248	141
Participants, n	309	176	405	345	246	44
Âge, années [1]	50.75 ± 15.95	47.50 ± 16.63	69.37 ± 9.91	69.75 ± 8.69	30.00 ± 8.21	31.37 ± 8.42
Hommes, %	44	48	35	45	40	53
Modèle de scanner	Philips Intera	Philips Intera	Siemens Magnetom TrioTim	Siemens BioGraph mMR PET-MR	Siemens Magnetom TrioTim	Siemens Magnetom TrioTim
Intensité du champ, Tesla	1.5	3	3	3	3	3
TR, ms [2]	9.81	9.60	2400	2300 (423) 2400 (30)	1900 (184) 2600 (64)	2400
TE, ms [2]	4.60	4.60	3.16	2.95 (423) 2.13 (30)	2.52 (184) 3.02 (64)	3.16
Résolution, mm^3 [2]	0.9 × 0.9 × 1.2	0.9 × 0.9 × 1.2	1.0 × 1.0 × 1.0	1.2 × 1.1 × 1.1 (423) 1.0 × 1.0 × 1.0 (30)	1.0 × 1.0 × 1.0	1.0 × 1.0 × 1.0
Dimensions, voxels [2]	256 × 256 × 150	256 × 256 × 150	176 × 256 × 256	176 × 240 × 256 (423) 176 × 256 × 256 (30)	176 × 256 × 256	176 × 256 × 256

[1] L'âge est exprimé comme moyenne ± écart-type.
[2] Le nombre d'images IRM avec les paramètres correspondant est indiqué entre parenthèses.

La sélection des datasets IRM est décrite en détail dans l'annexe 7.1.

4.2.1.2. Jeu de données de sujets voyageurs

Nous avons sélectionné 75 participants sains (CDR=0) de l'étude OASIS-3 qui avaient été scannés avec les deux scanners de Site3 et Site4 dans un intervalle de trois mois pour créer un jeu de données de sujets voyageurs pour une évaluation supervisée de variabilités liées au site. Ces sujets n'étaient pas inclus dans le jeu de données indépendant précédemment défini (section 4.2.1.1).

4.2.2. Prétraitements IRM

Nous avons d'abord retiré le crâne des images cérébrales T1w (en utilisant l'outil volBrain (Manjón et Coupé 2016)) et avons corrigé les effets d'homogénéité de champ magnétique (en utilisant l'algorithme N4ITK (Tustison et al. 2010)). Ensuite, nous avons recalé linéairement les images IRM dans un espace MNI $1mm^3$ avec l'outil FSL-FLIRT (Jenkinson et al. 2002) (six degrés de liberté) et avons mis à l'échelle les intensités en mettant la médiane au sein de chaque cerveau à 500. Nous avons trouvé que la normalisation de la médiane était moins sensible aux valeurs aberrantes que la normalisation du maximum, qui est beaucoup utilisée quand on applique de l'apprentissage profond à des images IRM (section 2.3.1).

4.2.3. Procédure d'harmonisation

4.2.3.1. Cadre et architectures

Le modèle développé est basé sur une configuration CycleGAN (Zhu et al. 2017). Etant donné deux domaines d'images, il apprend à *traduire* une image d'un domaine vers l'autre. L'entraînement a besoin d'un ensemble d'images de chaque domaine mais ne requiert pas de vérités terrain (i.e. les deux ensembles ne sont pas appariés). Avec un jeu de données IRM multicentrique, un site est défini comme le domaine de référence et les images des autres sites sont traduites vers le domaine de référence (les images IRM du domaine de référence ne sont pas modifiées).

Nous avons adapté le générateur U-Net et le discriminateur patchGAN décrits par Isola et al. (2017) avec des convolutions 3D pour le prétraitement d'images IRM 3D complètes (Figure 26). Le discriminateur patchGAN a un champ de réception de 38^3, que nous avons trouvé être un bon compromis entre l'accès à des informations contextuelles et la focalisation sur des détails d'imagerie locaux. Inspirés par Alami Mejjati et al. (2018), nous avons introduit l'application du masque original du cerveau après chaque génération d'image pour préserver la structure du cerveau et éviter que les voxels d'arrière-plan n'influencent l'entraînement. La fonction d'activation finale du générateur est linéaire pour les entrées supérieures à 0 (au lieu de *tanh*) pour permettre la génération de valeurs supérieures à 1.

Les deux générateurs et les deux discriminateurs traitent les volumes complets de dimensions 192^3.

4.2.3.2. Entraînement du modèle

Nous avons implémenté plusieurs stratégies pour accroître la stabilité de l'entraînement et la robustesse du modèle. Premièrement, les générateurs sont pré-entraînés à répliquer les images d'entrée avec tous les jeux de données indépendants et une fonction de coût L1. Deuxièmement, en vue de rendre le début de l'entraînement moins sensible à la variance,

un décroissement linéaire de 200 à 100 pour l'hyperparamètre pondérant la contrainte du cycle est appliqué (Wang et Lin 2018). Troisièmement, alors que la taille du batch est à 1 pour les générateurs, les discriminateurs s'entraînent avec un historique de 50 images générées (Shrivastava et al. 2017) et à chaque époque, chacun d'eux est mis à jour avec 4 images réelles, 2 images nouvellement générées et 2 images anciennement générées.

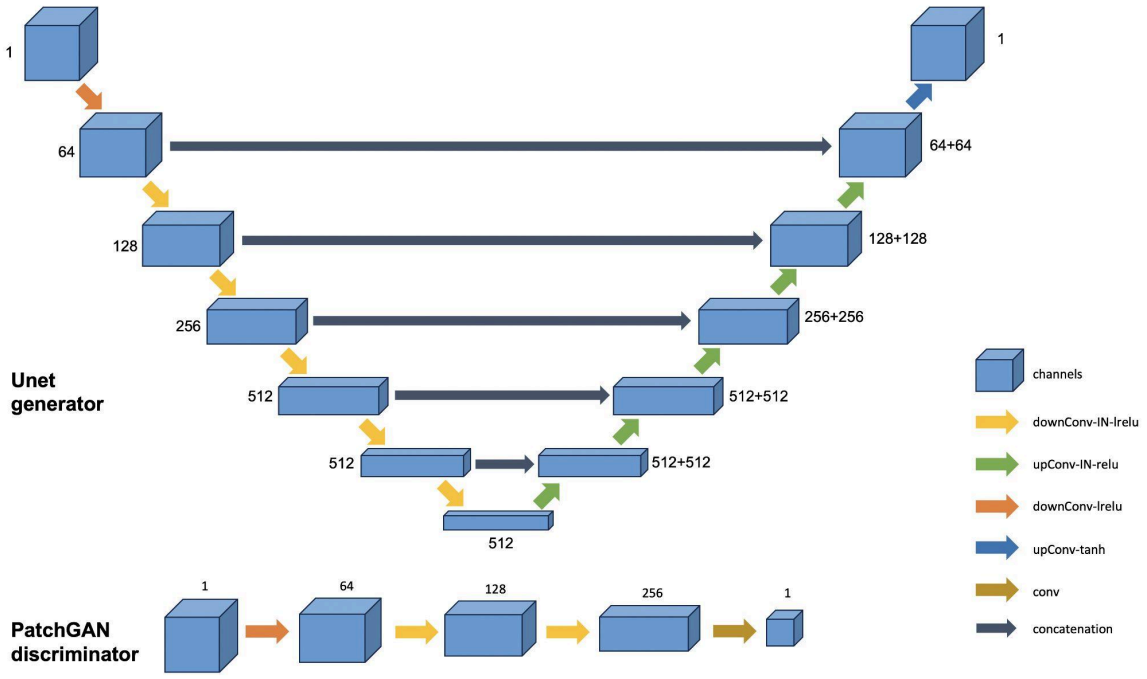


Figure 26 : Architectures des réseaux de la méthode CycleGAN proposée. Le nombre à côté de chaque boîte indique le nombre de canaux. Tous les noyaux de convolution sont 4^3 exceptés ceux de la dernière couche du discriminateur qui sont 3^3. downConv : convolution avec un pas de 2^3 ; upConv : convolution transposée avec un pas de 2^3 ; IN : normalisation d'instance ; lrelu : leaky relu avec une pente de 0.2.

Afin de sauvegarder de la mémoire GPU et d'accélérer les calculs, une politique de précision mixte (Micikevicius et al. 2018) est utilisée. De manière similaire à Wu et al. (2019), l'écart-type est remplacé par l'écart absolu moyen dans les normalisations d'instance (Ulyanov et al. 2017).

Plus de détails sur la procédure d'entraînement sont donnés dans l'annexe 7.2.

4.2.4. Harmonisations sites-appariés / multisite

Nous avons utilisé les six jeux de données indépendants (section 4.2.1.1) pour les expériences décrites dans cette section.

4.2.4.1. Configuration

Dans un premier ensemble d'expériences "sites-appariés", nous avons harmonisé trois paires de jeux de données : Site1 vs Site2, Site3 vs Site4 et Site5 vs Site6. Dans chaque paire, les deux sites ont des distributions d'âge similaires (vérification statistique dans l'annexe 7.3) ; cela évite les confusions entre des effets de site et des variabilités biologiques durant les phases d'entraînement et d'évaluation. Nous avons choisi les trois

paires afin de couvrir divers effets de site potentiels : une différence d'intensité du champ entre Site1 et Site2, des différents scanners IRM entre Site3 et Site4 et des différences de paramètre d'acquisition entre Site5 et Site6 (Tableau 4).

Nous avons également réalisé une harmonisation globale (multisite) dans laquelle cinq sites ont été harmonisés vers un site de référence. Bien que les images de Site1 aient été acquises avec un scanner 1.5T, nous l'avons choisi comme site de référence parce qu'il contient un nombre relativement grand d'images IRM et présente une gamme d'âge étendue (de 20 à 86 ans). Nous avons ainsi entraîné cinq modèles : Site1 vs Site2, Site1 vs Site3, … et Site1 vs Site6. Pour éviter de corriger des effets d'âge à cause de différences dans les distributions d'âge des jeux de données (annexe 7.4), nous avons divisé les images IRM en plusieurs tranches d'âge pour chaque entraînement et avons assigné une probabilité d'échantillonnage pour chaque tranche d'âge ; ainsi, à chaque étape, les deux ensembles ont la même probabilité d'échantillonnage pour chaque tranche d'âge. Pour éviter de sous-échantillonner les plus petites populations, nous avons développé un algorithme qui, à partir de deux jeux d'entraînement et des tranches d'âge, donne une probabilité d'échantillonnage pour chacune de ces tranches. L'algorithme est décrit dans l'annexe 7.5 avec les probabilités d'échantillonnage calculées et utilisées pour les différentes harmonisations vs Site1. Il doit être noté que nous n'avons pas adopté cette stratégie d'échantillonnage pour l'harmonisation avec Site2, étant donné que les deux distributions d'âges sont similaires (même modèle utilisé que dans l'expérience sites-appariés correspondante).

4.2.4.2. Évaluations quantitatives

Nous avons mesuré diverses caractéristiques cérébrales avant et après harmonisation pour analyser les variabilités inter-sites et les informations biologiques. Plus spécifiquement, nous avons utilisé l'outil FSL-FAST (Zhang et al. 2001) pour segmenter la MG, la MB et le LCS.

4.2.4.2.1. Variabilités inter-sites de caractéristiques IRM

Pour une analyse quantitative de différences inter-sites, nous avons défini trois groupes de caractéristiques IRM : volumes de tissus, IQMs et radiomiques de premier ordre. Les volumes de tissus et les IQMs sont listés dans le Tableau 5 et sont basés sur ceux décrits par Esteban et al. (2017). Les 36 radiomiques de premier ordre sont extraites avec l'outil PyRadiomics (van Griethuysen et al. 2017) en utilisant les masques de MG et de MB. Pour limiter l'effet des caractéristiques redondantes et bruitées, nous avons appliqué une standardisation suivie d'une ACP pour projeter les radiomiques dans un espace 2D et visualiser des clusters de site potentiels.

Étant donné que ces caractéristiques peuvent être influencées par l'âge, des hétérogénéités ne devraient pas être supprimées avec l'harmonisation. Dans l'expérience multisite, nous les avons donc étudiées sur plusieurs tranches d'âge spécifiques (20-30, 50-60 et 60-70) séparément. Pour chacune, nous avons uniquement inclus les sites avec plus de 20 participants.

4.2.4.2.2. Prédiction d'âge

La prédiction d'âge consiste en l'entraînement d'un modèle d'apprentissage automatique pour la prédiction de l'âge d'un individu à partir de données d'IRM cérébrale et a été largement investiguée ces dernières années (Sajedi et Pardakhti 2019). Des méthodes

d'apprentissage profond peuvent faire des estimations précises de l'âge en traitant directement les images (Cole et al. 2017; Gautherot et al. 2021). Toutefois, la généralisation de ce type de modèles en imagerie médicale peut être difficile (Zech et al. 2018). Dans cette étude, nous avons implémenté un modèle de prédiction d'âge similaire à celui décrit par Cole et al. (2017) ; le but était d'évaluer la capacité de l'approche d'harmonisation à conserver ou à accentuer des motifs biologiques.

Tableau 5 : Volumes de tissus et métriques de qualité d'image pour la section 4.

Volumes de tissus	
icv_gm	Les volumes intracrâniens de MG, MB et LCS, chacun divisé par le volume intracrânien total.
icv_wm	
icv_csf	
Métriques de qualité d'images	
cjv	Un coefficient de variation jointe entre les intensités de MG et MB.
efc	L'entropie de Shannon des intensités de voxels.
snr_gm	Un rapport signal sur bruit pour la MG, la MB et le LCS.
snr_wm	
snr_csf	
wm2max	Le ratio entre l'intensité médiane de MB et le 95e percentile de l'intensité globale du cerveau.
rpve_gm	Une estimation des effets de volumes partiels résiduels pour la MG, la MB et le LCS.
rpve_wm	
rpve_csf	
fwhm	Largeur à mi-hauteur : une estimation de la netteté de l'image utilisant AFNI *3dFWHMx*.

MG : matière grise ; MB : matière blanche ; LCS : liquide cérébrospinal

Pour chaque expérience sites-appariés, nous avons sélectionné le site avec le plus grand nombre d'images IRM comme référence de l'harmonisation (i.e. Site1, Site3 et Site5), entraîné un modèle de prédiction d'âge que nous avons ensuite évalué sur les images des autres sites (i.e. Site2, Site4 et Site6) avant et après harmonisation. En vue de mesurer des performances de référence, nous avons divisé aléatoirement les images IRM des trois sites de référence en jeux d'entraînement et de test avec une stratification des âges ; cela a résulté en 50, 117 et 30 images IRM de test pour Site1, Site3 et Site5 respectivement.

Pour l'expérience multisite, nous avons évalué le modèle entraîné sur des images IRM de Site1 (l'ensemble de référence) sur les cinq autres jeux de données avant et après harmonisation. Nous avons également mis en place un modèle de prédiction d'âge pour évaluer l'intérêt de l'harmonisation sur un jeu d'entraînement multicentrique de grande taille. Nous avons pour cela divisé aléatoirement nos données en un jeu d'entraînement de 1863 images et un jeu de test de 448 images (aucun sujet présent dans les deux jeux), tout en conservant la proportion d'images au sein de chaque site. Ensuite, nous avons entraîné et

évalué deux modèles de prédiction d'âge : un sans harmonisation et l'autre avec les images harmonisées vers Site1.

Pour analyser les prédictions d'âge, nous avons calculé l'erreur absolue moyenne (EAM) et la différence d'âge prédit moyenne (DAPM, la moyenne des âges réels soustraits aux âges prédits). Pour rendre compte de la régression vers la moyenne d'entraînement (RME) (Butler et al. 2021) dans l'expérience multisite avec le modèle de prédiction d'âge entraîné sur les images IRM de Site1, nous avons également calculé la déviation par rapport à la moyenne d'entraînement (DME) en soustrayant l'âge moyen du jeu de test de celui du jeu d'entraînement. Les résultats sont indiqués en années dans ce manuscrit.

Des détails sur les étapes d'entraînement pour les modèles de prédiction d'âge sont donnés dans l'annexe 7.6.

4.2.4.2.3. Corrélation entre l'âge et le volume de matière grise

L'une des principales caractéristiques du vieillissement cérébral chez les adultes est une réduction constante du volume de MG (Ge et al. 2002; Hedman et al. 2012; Watanabe et al. 2013). En vue d'évaluer l'effet de l'harmonisation sur ce motif, nous avons calculé le coefficient de corrélation de Pearson entre le volume de MG et l'âge.

4.2.4.2.4. Validations radiologiques

Afin d'évaluer la conservation de motifs radiologiques après harmonisation, un sous-ensemble des images T1w des jeux de données indépendants a été revu par un neuroradiologue certifié (GK). L'atrophie corticale globale (ACG), l'atrophie temporale médiale (ATM), les espaces périvasculaires dilatés (EPDs) et la taille des ventricules ont été estimés étant donné que ces caractéristiques sont associées à un vieillissement normal et/ou des troubles liées à l'âge (comme la maladie d'Alzheimer, la maladie des petits vaisseaux ou l'hydrocéphalie à pression normale). L'ACG a été cotée sur une échelle semi-quantitative à 4 points adaptée de celle décrite par Pasquier et al. (1996) (0 = absente, 1 = légère, 2 = modérée, 3 = sévère) et l'ATM a été cotée sur une échelle semi-quantitative à 5 points (Scheltens et al. 1995). Les EPDs ont été identifiés comme de petites intensités linéaires nettement délimitées de LCS (ou des structures proches des intensités de LCS) mesurant < 3 mm suivant le trajet des vaisseaux perforants ou médullaires (Wardlaw et al. 2013). Le nombre d'EPDs dans le ganglion de la base (EPD-GB, sur la première coupe au-dessus de la commissure antérieure) et dans le centre semi-ovale (EPD-CS, sur la première coupe au-dessus des ventricules latéraux) a été coté comme suit : 0 = pas d'EPD, 1 = 1 à 9 EPDs, 2 = 10 à 20 EPDs, 3 = 21 à 40 EPDs et 4 = 40 ou plus EPDs (MacLullich et al. 2004). L'index d'Evans a été déterminé par la largeur transversale maximale de la corne frontale du ventricule latéral, perpendiculaire à la ligne médiane sagittale, sur une section axiale 2D parallèle au plan commissure antérieure - commissure postérieure, divisée par la largeur transversale maximale de la cavité intracrânienne sur le même plan (Miskin et al. 2017).

Pour les expériences sites-appariés, 10 participants ont été sélectionnés aléatoirement dans Site2, Site4 et Site5 et une image de chacun a ensuite été cotée avant et après harmonisation vers Site1, Site3 et Site6, respectivement. Pour l'expérience multisite, 6 participants ont été sélectionnés aléatoirement dans Site2, Site3, … et Site6 et une image de chacun a ensuite été cotée avant et après harmonisation vers Site1. Les 120 images IRM cotées ont été mélangées et anonymisées avant la revue. Nous avons quantifié la consistance des scores avec l'harmonisation en calculant le kappa pondéré quadratique

pour les mesures ordinales (i.e. ACG, ATM, EPD-GB et EPD-CS) et une corrélation intraclasse (Fisher 1992) pour l'index d'Evans. Nous avons interprété la consistance des cotations à partir des kappa et des corrélations intraclasse de la manière suivante : pauvres en dessous de 0.40, justes entre 0.40 et 0.59, bonnes entre 0.60 et 0.74 et excellentes au-dessus de 0.74 (Cicchetti 1994).

4.2.4.2.5. Inférences statistiques

Nous avons comparé les volumes de tissus et les IQMs des sites en utilisant des t-tests bilatéraux pour les expériences sites-appariés et une ANOVA à un facteur pour l'expérience multisite. Pour les résultats de prédiction d'âge, nous avons utilisé des tests de Wilcoxon des rangs signés bilatéraux pour comparer les erreurs de prédiction avant et après harmonisation. Nous avons comparé les coefficients de Pearson en utilisant le test de Steiger bilatéral (Steiger 1980). Avec la procédure de Benjamini-Hochberg, nous avons corrigé les valeurs p (i) pour chaque comparaison de volumes de tissus et de IQMs, (ii) pour les trois comparaisons de prédiction d'âge dans les expériences sites-appariés et (iii) pour les cinq comparaisons de prédiction d'âge dans l'expérience multisite avec le modèle entraîné sur les images IRM de Site1. Pour favoriser une indépendance entre les échantillons, nous avons moyenné les données pour chaque participant avant chaque test inférentiel. Nous avons fixé le seuil de significativité statistique à 0.05 pour les valeurs p.

4.2.5. Harmonisation sur des sujets voyageurs

Nous avons utilisé notre jeu de données de sujets voyageurs (section 4.2.1.2) pour évaluer la faisabilité de notre modèle à transformer des images d'un site vers leur équivalent dans l'autre. Nous avons réutilisé les modèles d'harmonisation précédemment entraînés pour les expériences sites-appariés (Site3 vs Site4, section 4.2.4.1) pour harmoniser les 76 sujets voyageurs dans les deux directions. Nous avons calculé le SSIM (Wang et al. 2004) pour chaque paire d'images avec une gamme dynamique fixée à 1000. Avant de calculer le SSIM, nous avons supprimé les coupes d'arrière-plan.

Nous avons comparé notre approche 3D avec un CycleGAN 2D adapté de Zhu et al. (2017). Pour exploiter les trois orientations, nous avons entraîné trois modèles 2D et généré les volumes de sortie finaux avec une inférence 2.5D (Cackowski et al. 2023; Dewey et al. 2019). Pour l'entraînement, nous avons suivi une approche précédemment proposée consistant en l'utilisation exclusive des coupes qui contiennent au moins 1% de pixels différents de 0 (Bashyam et al. 2022; Cackowski et al. 2023). Plus de détails sur le CycleGAN 2D implémenté sont donnés dans l'annexe 7.7.

Pour chaque harmonisation, nous avons effectué des tests de Wilcoxon des rangs signés bilatéraux pour comparer les SSIMs avant et après harmonisation. Nous avons comparé de la même manière les SSIMs obtenus avec les deux méthodes CycleGAN. Nous avons corrigé les valeurs p avec la procédure Benjamini Hochberg.

4.2.6. Accès aux données et au code

Toutes les images IRM viennent de bases de données publiques (section 4.2.1). Les codes Python pour le modèle d'harmonisation, la prédiction d'âge, l'extraction des IQMs et des volumes de tissus et la fonction utilisée pour équilibrer les distributions d'âge sont accessibles publiquement : https://gitlab.com/RocaV/3d_cyclegan_mri_harmonization.

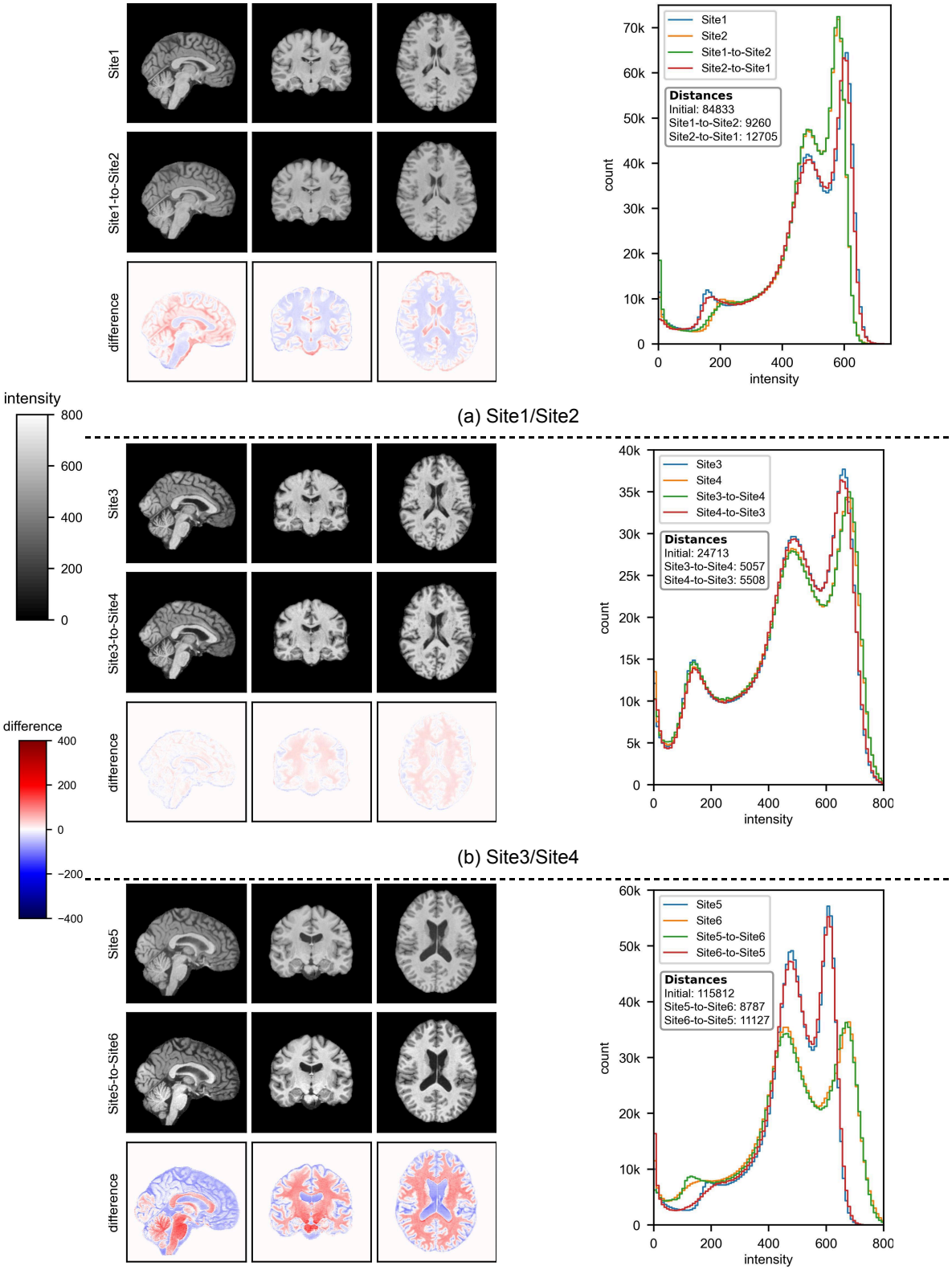


Figure 27 : Coupes d'images et histogrammes moyens d'intensités cérébrales dans les expériences sites-appariés de la section 4. Une image IRM a été sélectionnée aléatoirement pour chaque harmonisation illustrée. Les différences correspondent à une soustraction par voxel de l'image originale à l'image harmonisée. 100 tranches d'âge consécutives sont définies pour les histogrammes moyens. Des distances euclidiennes avec et sans harmonisation entre les sites *source* et *cible* sont indiquées.

4.3. Résultats

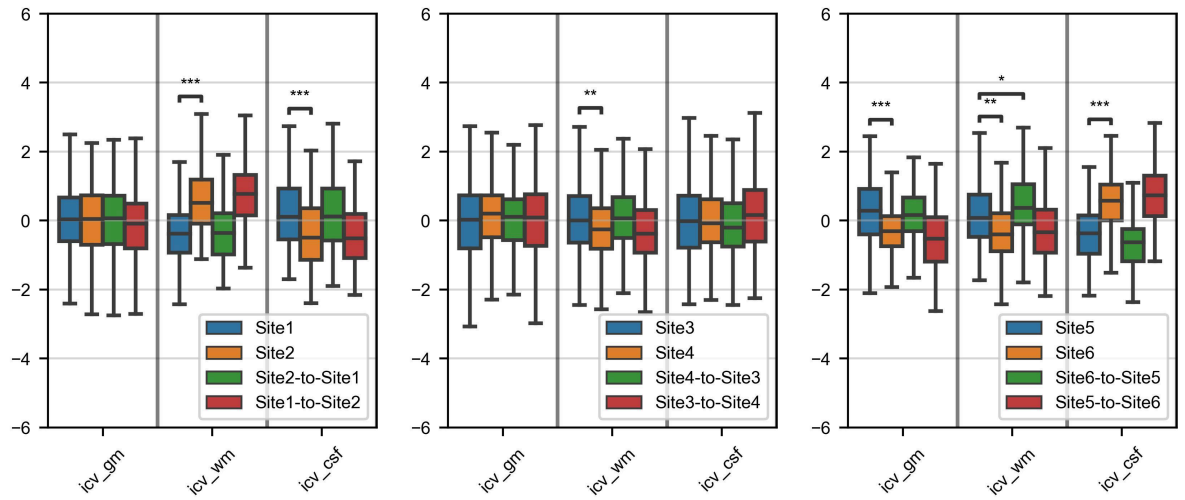
4.3.1. Expériences sites-appariés

4.3.1.1. Comparaison d'images et d'histogrammes

Chaque paire de sites présentait plus ou moins de variabilités sur les distributions d'intensités de voxel (Figure 27). La différence de contraste était significative entre Site5 et Site6 mais elle a été bien corrigée par l'harmonisation (Figure 27c). Les différences étaient plus subtiles pour Site1/Site2 et Site3/Site4 mais l'harmonisation a quand même permis une homogénéisation des distributions (Figures 27a et 27b). Les six procédures d'harmonisation ont décru la distance Euclidienne (Cha 2008) entre les histogrammes moyens d'intensités cérébrales de 85.69% en moyenne.

4.3.1.2. Variabilités inter-sites de caractéristiques IRM

Les volumes cérébraux étaient significativement associés au site pour les paires Site1/Site2 et Site5/Site6 (Figures 28a et 28c, respectivement). Pour la paire Site3/Site4, seuls les volumes de MB ont montré une variabilité inter-sites significative (Figure 28b). Après harmonisation, les variabilités inter-sites étaient plus faibles et seule la différence de volume de MB entre Site5 et Site6 était toujours statistiquement significative (p=0.0461).

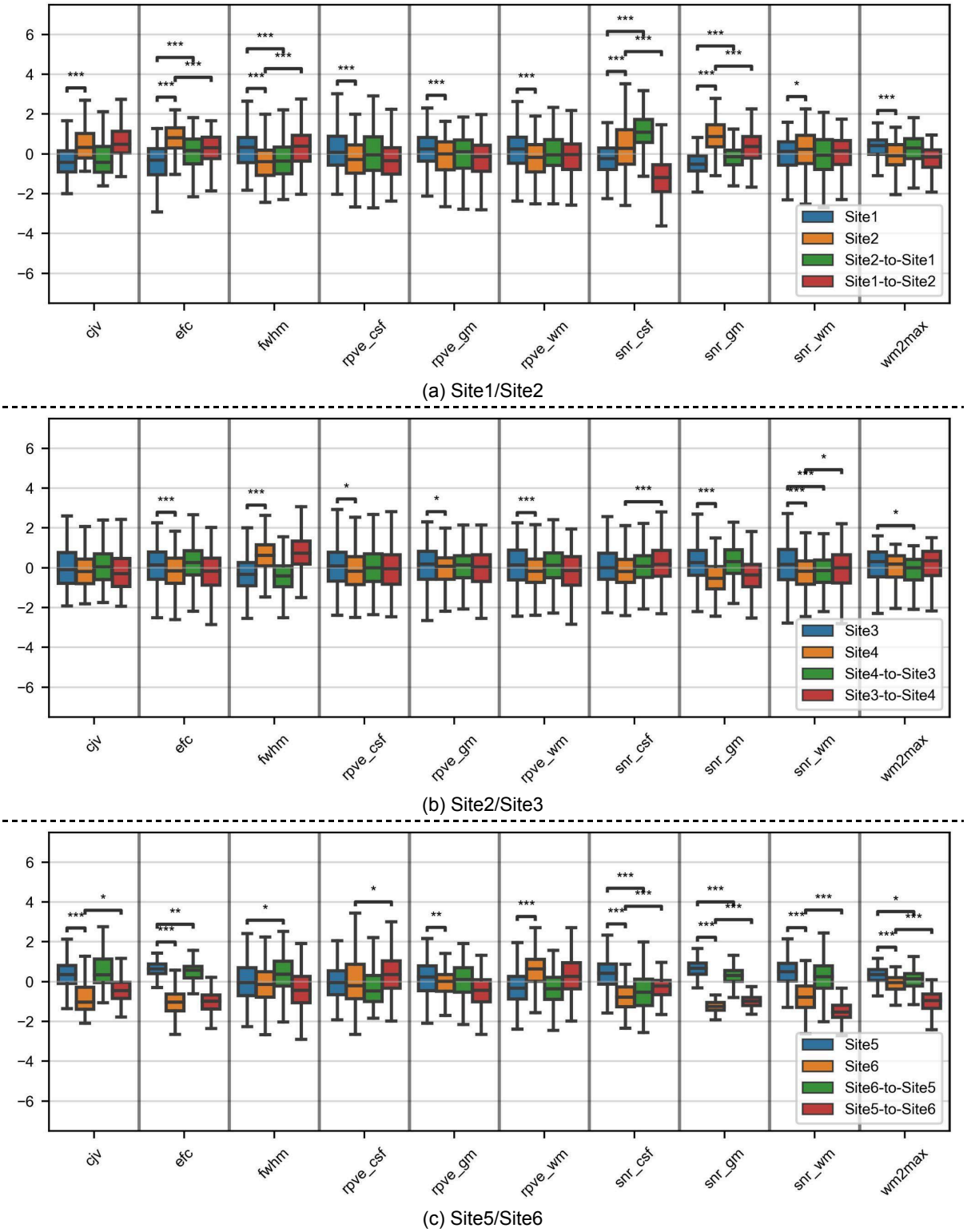


(a) Site1/Site2 (b) Site3/Site4 (c) Site5/Site6

Figure 28 : Distributions des volumes de tissus dans les expériences sites-appariés de la section 4. Les volumes sont divisés par le volume intracrânien total. Pour chaque sous-figure et chaque tissu, l'axe des Y est un Z-score basé sur les deux ensembles d'images originales. Des astérisques indiquent des t-tests significatifs avant et après harmonisation entre les sites *source* et *cible* (* : p<0.05; ** : p<0.01; *** : p<0.001).

Une majorité des IQMs présentaient des effets de site dans les trois expériences avant harmonisation (Figure 29). Ces effets étaient particulièrement significatifs pour les paires Site1/Site2 et Site5/Site6 (Figures 29a et 29c, respectivement). Pour la plupart des IQMs (e.g. *cjv* et les effets de volumes partiels), l'harmonisation a été effective dans les deux directions - particulièrement pour la paire Site3/Site4 (Figure 29b) - mais pas pour *snr_csf*,

pour laquelle l'harmonisation a mené à des changements chaotiques dans la plupart des expériences.



(a) Site1/Site2

(b) Site2/Site3

(c) Site5/Site6

Figure 29 : Distributions des métriques de qualité d'image dans les expériences sites-appariés de la section 4. Pour chaque sous-figure et chaque métrique, l'axe des Y est un Z-score basé sur les deux ensembles d'images originales. Des astérisques indiquent des t-tests significatifs avant et après harmonisation entre les sites *source* et *cible* (* : p<0.05; ** : p<0.01; *** : p<0.001).

Dans la Figure 30, nous présentons les résultats de l'ACP sur les radiomiques. La dissociation était claire pour la paire Site5/Site6 (Figure 30g) et après harmonisation, les échantillons étaient plus mixés même si des clusters de site étaient toujours distinguables (Figures 30h et 30i). L'effet de site était moins clair mais significatif quand même pour Site1 et Site2 (Figure 30a) et a également été réduite avec succès avec l'harmonisation (Figures 30b et 30c). Même si les échantillons de Site3 et Site4 étaient moins séparés à la base (Figure 30d), des chevauchements plus importants ont été obtenus également avec l'harmonisation (Figures 30e et 30f).

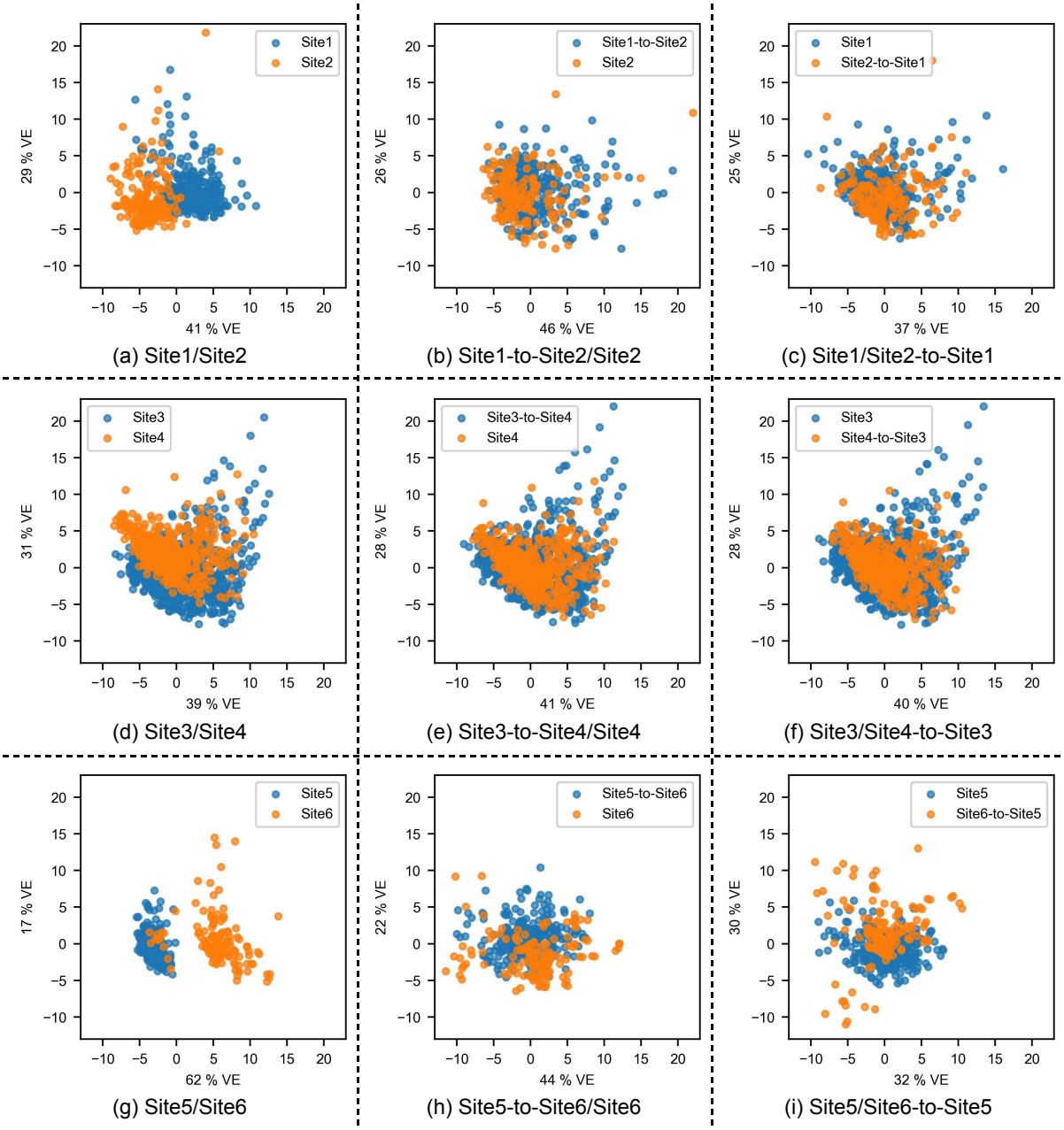
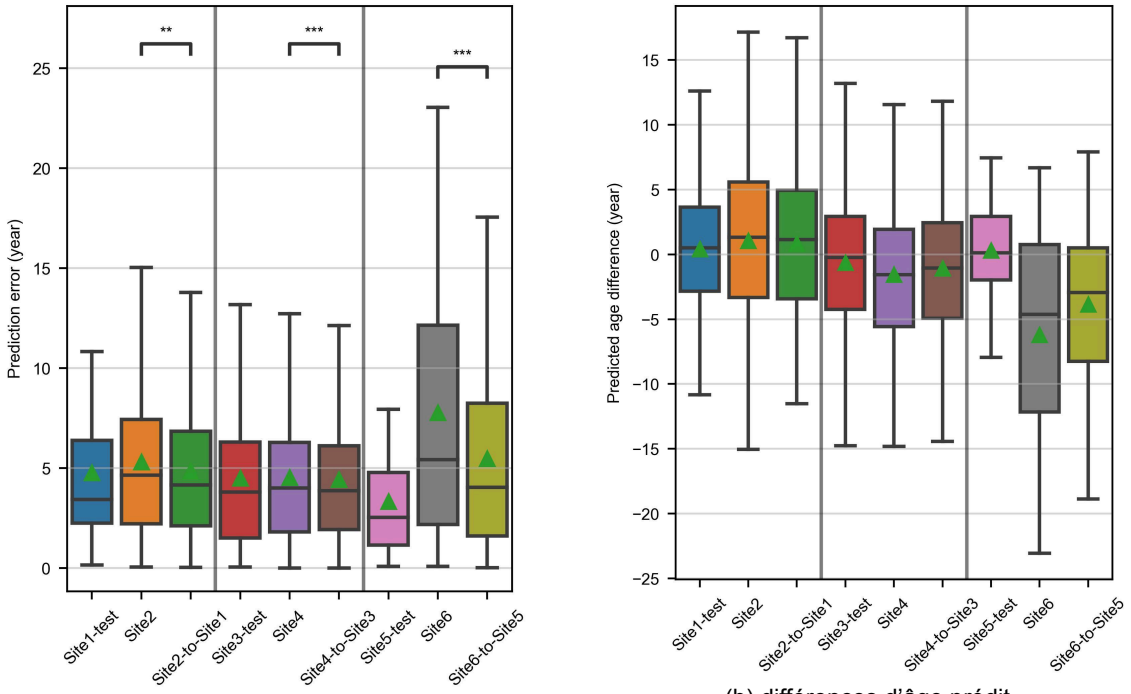


Figure 30 : Composantes de l'ACP sur les radiomiques dans les expériences sites-appariés de la section 4. Les axes des X et des Y correspondent respectivement aux premier et deuxième axes de l'ACP. VE : variance expliquée.

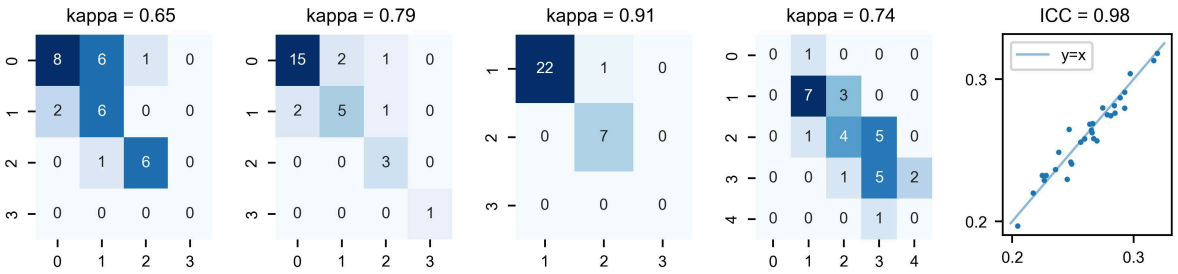
4.3.1.3. Prédiction d'âge



(a) erreurs de prédiction

(b) différences d'âge prédit

Figure 31 : Prédiction d'âge dans les expériences sites-appariés de la section 4. Dans **(a)**, les astérisques indiquent des tests de Wilcoxon significatifs (* : p<0.05; ** : p<0.01; *** : p<0.001). Dans **(b)**, la différence d'âge prédit correspond à l'âge prédit moins l'âge réel.



(a) atrophie corticale globale

(b) atrophie temporale médiale

(c) nombre d'espaces périvasculaires dilatés dans le ganglion de la base

(d) nombre d'espaces périvasculaires dilatés dans le centre semi-ovale

(e) index d'Evans

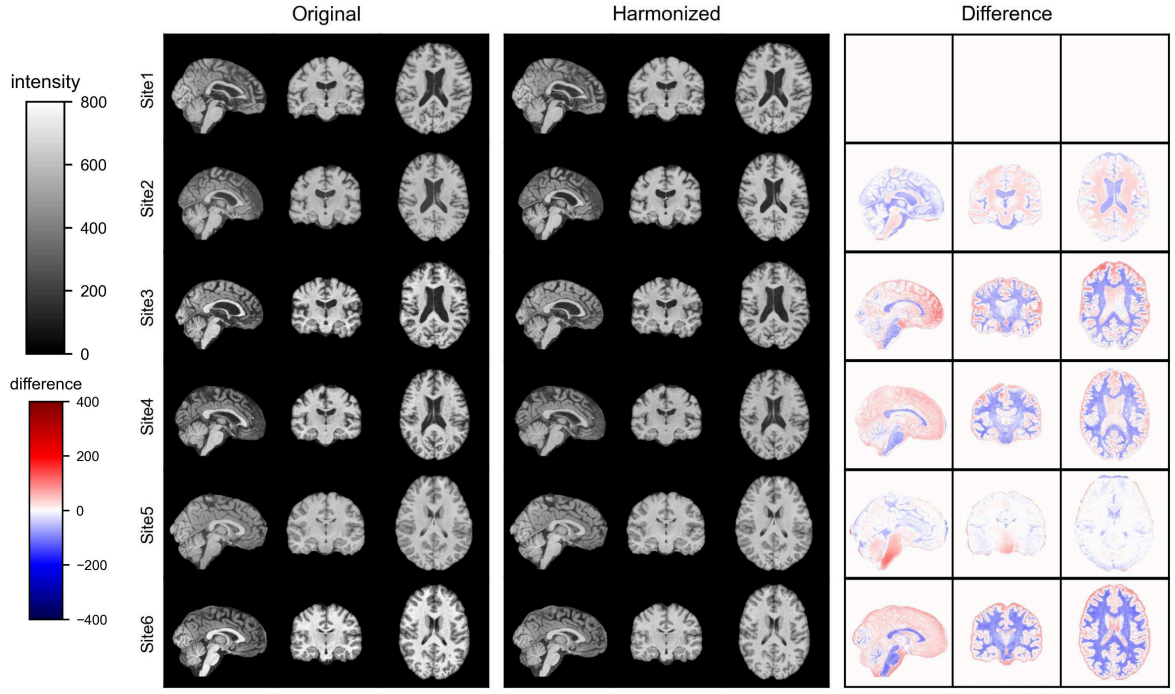
Figure 32 : Scores radiologiques dans les expériences sites-appariés de la section 4. Les résultats sont représentés dans un tableau de contingence pour les variables ordinales (la profondeur de couleur est proportionnelle à l'effectif dans chaque case). Les axes des Y et des X correspondent respectivement aux valeurs des images originales et harmonisées. ICC : corrélation intraclasse.

La Figure 31 illustre les résultats relatifs aux trois expériences de prédiction d'âge. Dans toutes les expériences, l'EAM du jeu de test était plus faible que celle du jeu de généralisation et l'harmonisation a résulté en une baisse significative des erreurs de prédiction - même si l'amplitude de la baisse était variable (Figure 31a). Pour les trois jeux de généralisation, la DAPM était plus proche de 0 après harmonisation (particulièrement

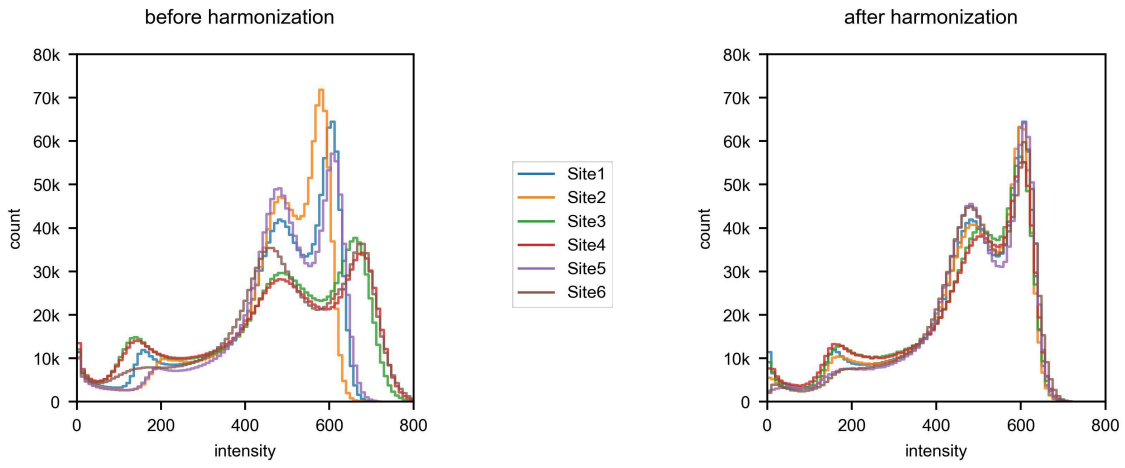
pour la paire Site5/Site6), ce qui signifie que les motifs de sur/sous-estimation causés par des effets de site ont été en partie corrigés (Figure 31b).

4.3.1.4. Scores radiologiques

La consistance des cotations radiologiques entre les images avant et après harmonisation était bonne pour l'ACG (Figure 32a) et l'EPS-CS (Figure 32d) et excellent pour l'ATM (Figure 32b), l'EPS-GB (Figure 32c) et l'index d'Evans (Figure 32e).



(a) exemples d'harmonisation vers Site1



(b) histogrammes moyens avant et après harmonisation

Figure 33 : Coupes d'images et histogrammes moyens d'intensités cérébrales dans l'expérience multisite de la section 4. Dans **(a)**, une image IRM a été sélectionnée aléatoirement pour chaque site et des coupes sont montrées avant et après harmonisation vers Site1. Les différences correspondent à une soustraction par voxel de l'image originale à l'image harmonisée. Dans **(b)**, 100 tranches d'âge consécutives sont définies pour les histogrammes moyens.

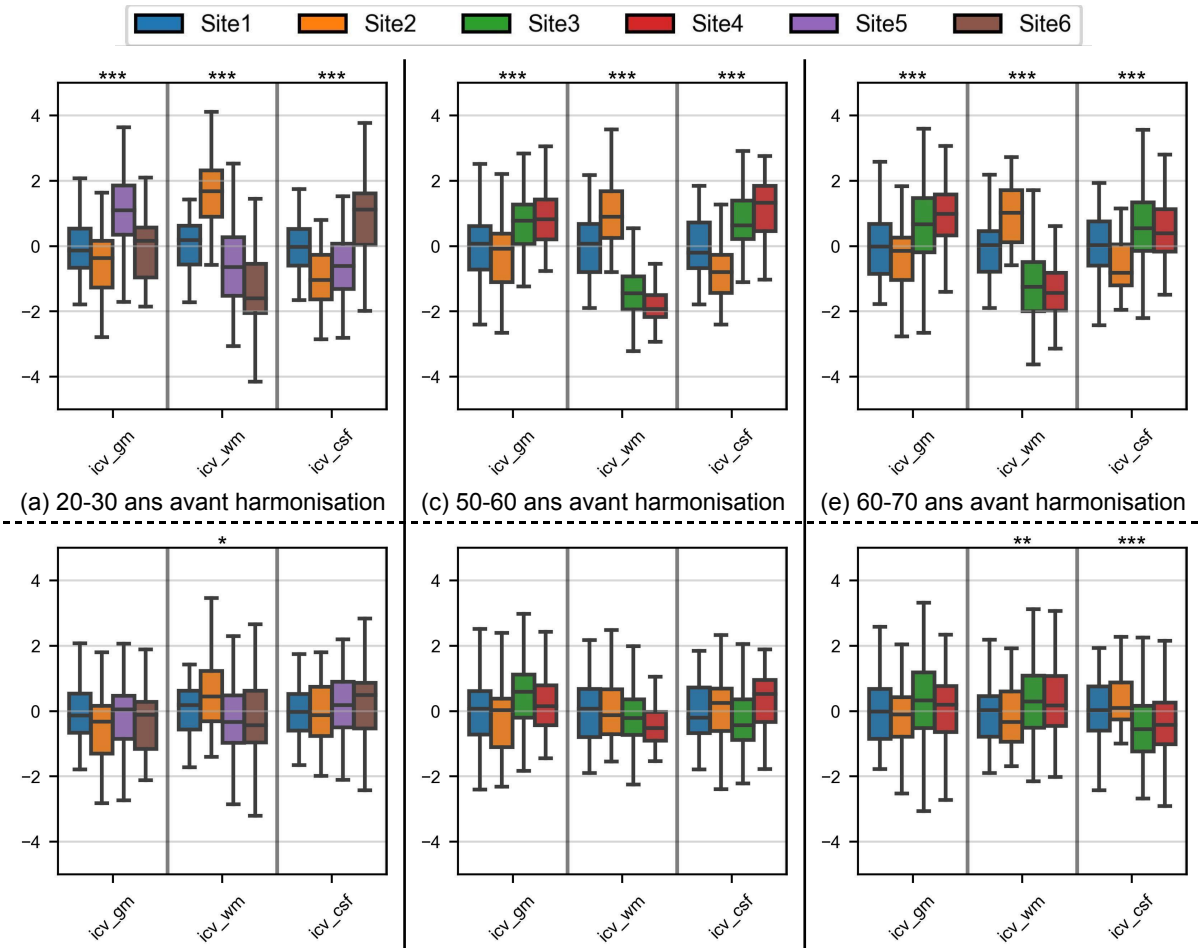
4.3.2. Expérience multisite

4.3.2.1. Comparaison d'images et d'histogrammes

La Figure 33a montre que, par rapport aux images de Site1, les images originales de Site3, Site4 et Site6 étaient plus lumineuses et avaient un contraste MG/MB élevé alors que les images de Site2 et Site5 étaient plus sombres. Ces variabilités inter-sites n'étaient plus visibles sur les images IRM harmonisées. Néanmoins, les structures cérébrales semblent avoir été conservées.

La Figure 33b présente les histogrammes moyens. En l'absence d'harmonisation, ils différaient d'un site à l'autre. Pour quantifier ces différences, nous avons calculé la somme des 100 écart-types des décomptes moyens de voxel par site associés aux 100 classes d'histogramme et obtenu une valeur de 465197. Les histogrammes étaient plus similaires après harmonisation et notre index d'hétérogénéité a été réduit à 135383 (une réduction de 70.90%).

4.3.2.2. Variabilités inter-sites de caractéristique IRM



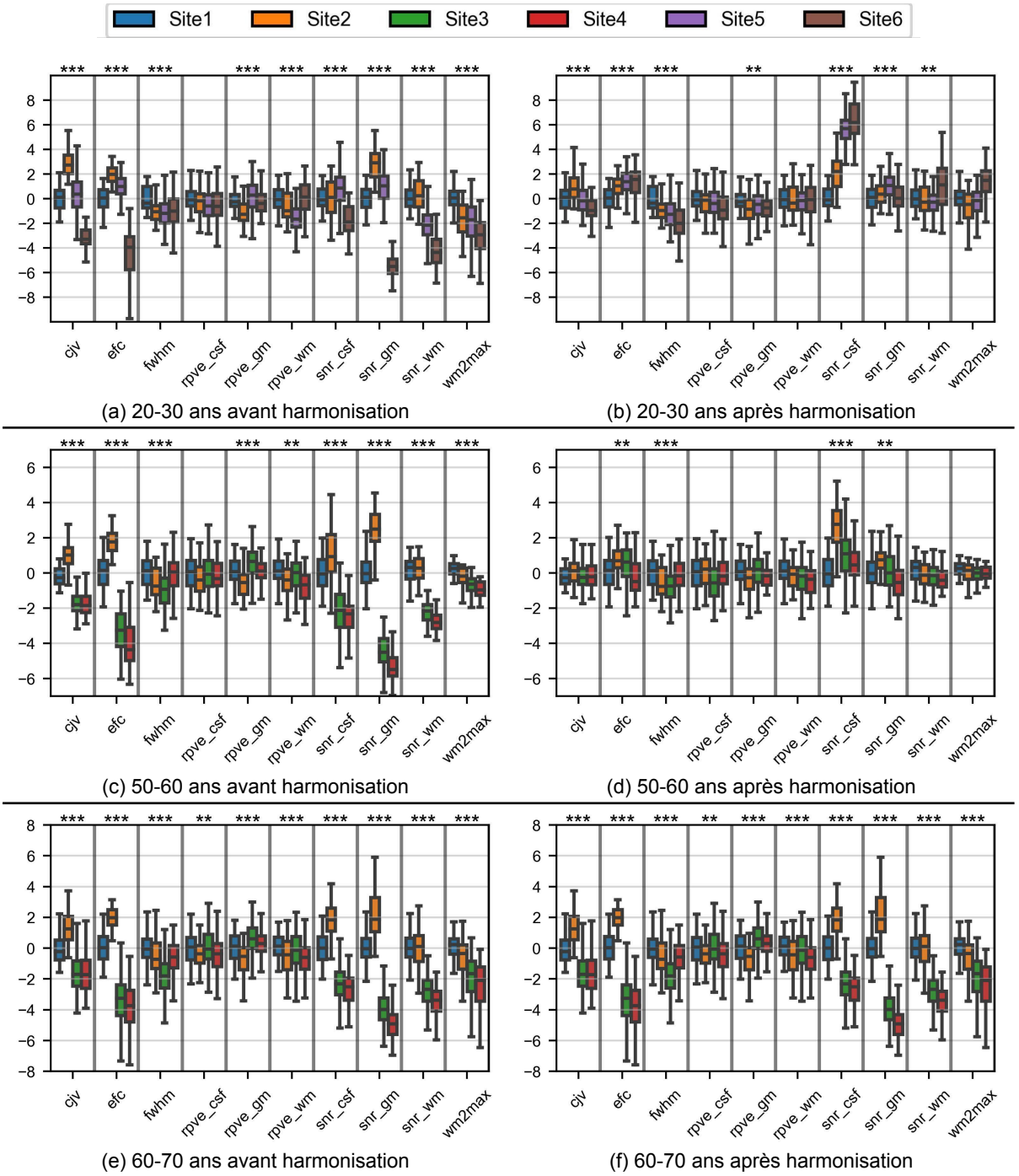
(a) 20-30 ans avant harmonisation (c) 50-60 ans avant harmonisation (e) 60-70 ans avant harmonisation

(b) 20-30 ans après harmonisation (d) 50-60 ans après harmonisation (f) 60-70 ans après harmonisation

Figure 34 : Distributions des volumes de tissus par tranche d'âge dans l'expérience multisite de la section 4. Les volumes sont divisés par le volume intracrânien total. Pour chaque sous-figure et chaque tissu, l'axe des Y est un Z-score basé sur les images de Site1 dans la tranche d'âge donnée. Des astérisques indiquent des tests ANOVA significatifs (* : p<0.05; ** : p<0.01; *** : p<0.001).

La mesure pour chaque volume de tissu dans chaque tranche d'âge variait significativement d'un jeu de données à l'autre (Figures 34a, 34c et 34e). La variabilité était plus faible après harmonisation (Figures 34b, 34d et 34f), excepté pour le volume de LCS chez les 60-70 ans (Figure 34f).

D'importantes différences inter-sites pour presque toutes les IQMs étaient apparentes sur les images IRM originales (Figures 35a, 35c et 35e). Globalement, ces différences ont été réduites par l'harmonisation excepté pour *fwhm* et *snr_csf*.



(a) 20-30 ans avant harmonisation

(b) 20-30 ans après harmonisation

(c) 50-60 ans avant harmonisation

(d) 50-60 ans après harmonisation

(e) 60-70 ans avant harmonisation

(f) 60-70 ans après harmonisation

Figure 35 : Distributions des métriques de qualité d'image par tranche d'âge dans l'expérience multisite de la section 4. Pour chaque sous-figure et chaque tissu, l'axe des Y est un Z-score basé sur les images de Site1 dans la tranche d'âge donnée. Des astérisques indiquent des tests ANOVA significatifs (* : p<0.05; ** : p<0.01; *** : p<0.001).

En se basant sur les composantes des ACPs appliquées aux radiomiques, des clusters de site étaient facilement distinguables sur les données originales (Figures 36a, 36c et 36e). Une exception était visible chez les 20-30 ans, où les composantes de Site1 se superposaient à celles de Site5 (Figure 36a). Les clusters étaient plus difficilement distinguables après harmonisation (Figures 36b, 36d et 36f).

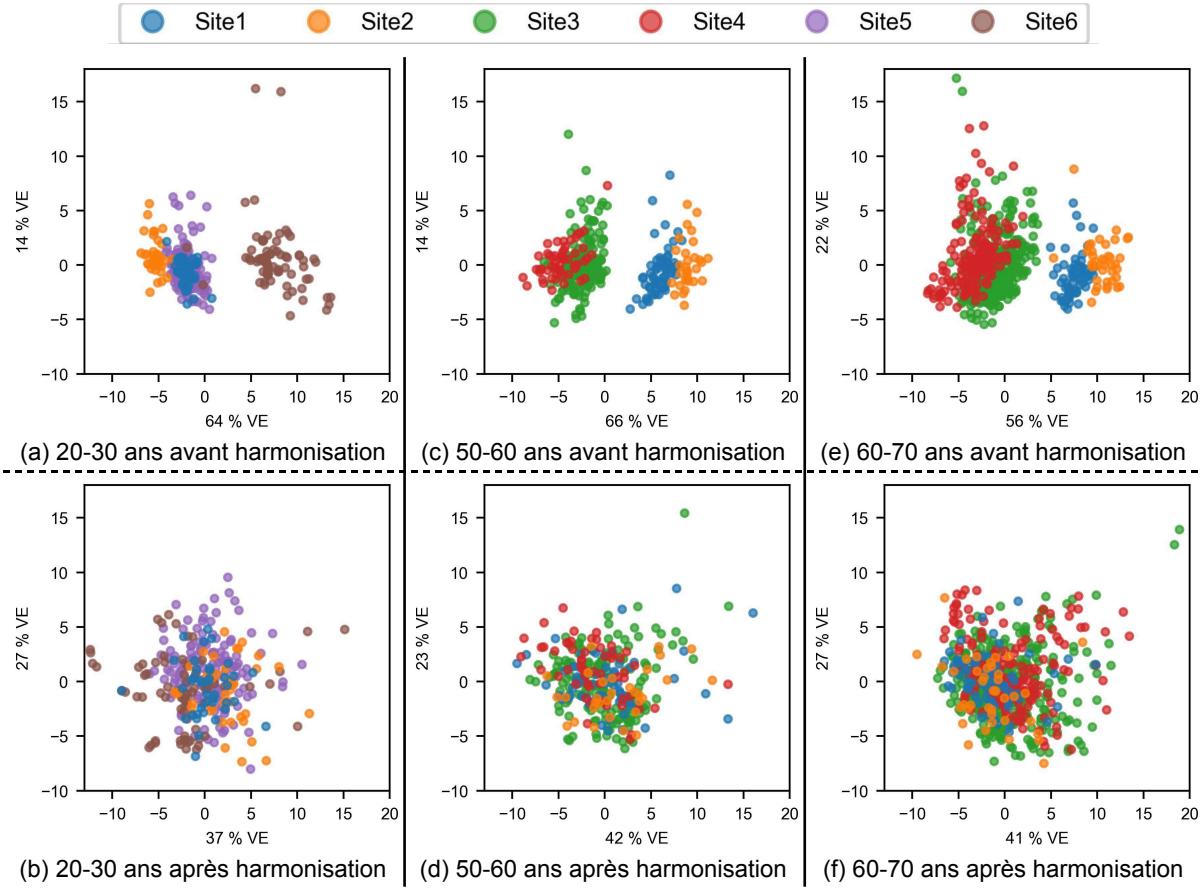
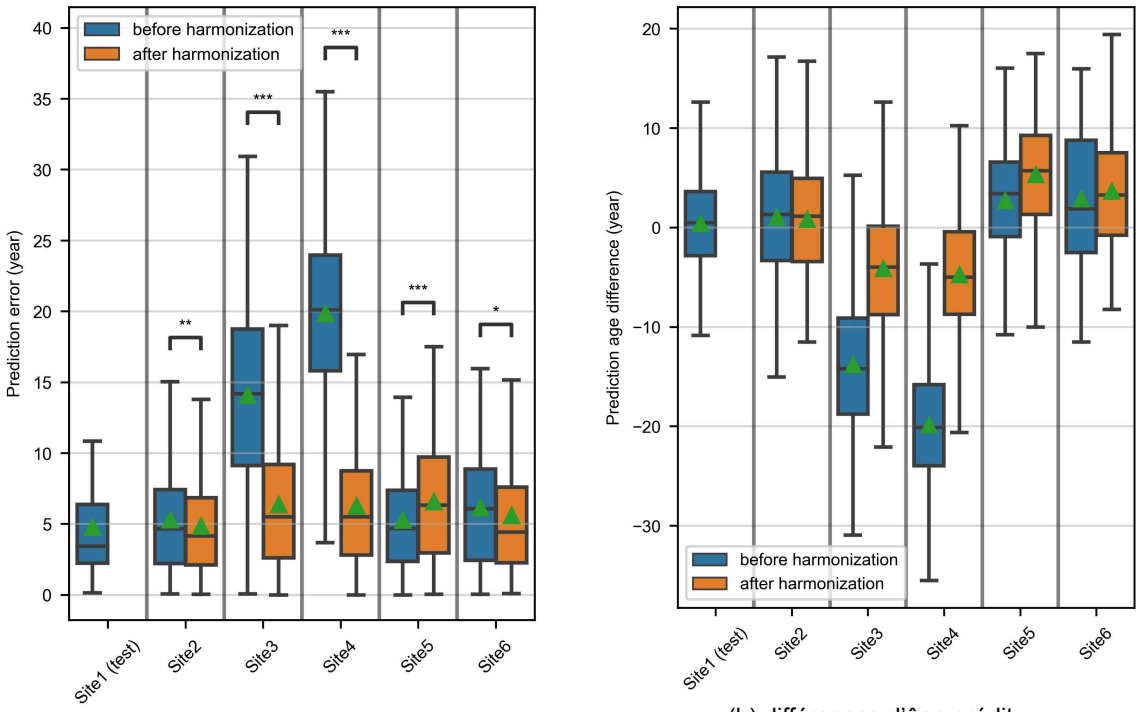


Figure 36 : Composantes de l'ACP sur les radiomiques dans l'expérience multisite de la section 4. Les axes des X et des Y correspondent respectivement aux premier et deuxième axes de l'ACP. VE : variance expliquée.

4.3.2.3. Prédiction d'âge

4.3.2.3.1. Jeu d'entraînement unicentrique

La Figure 37 illustre les résultats produits par le modèle de prédiction d'âge entraîné avec les images de Site1 et appliqué sur tous les jeux de données. La Figure 37a montre que, comme dans les expériences sites-appariés, l'EAM sur le jeu de test était plus faible que celles sur les jeux de généralisation. Après harmonisation, les erreurs de prédiction ont réduit significativement pour Site2, Site3, Site4 et Site6. À l'inverse, elles ont augmenté significativement pour SIte5. On peut voir dans la Figure 37b que, après harmonisation, la DAPM était plus proche de 0 pour Site2, Site3 et Site4 mais plus éloignée pour Site5 et Site6 ; cela suggère que les motifs de sous/sur-estimation liés au site ont été atténués dans le premier cas et accentués dans le second. Pour chaque jeu de données autre que Site1, la DAPM était du même signe que la DME, ce qui est un indicateur d'une RME.



(a) erreurs de prédiction

(b) différences d'âge prédit

Figure 37 : Prédiction d'âge du modèle entraîné sur Site1 dans l'expérience multisite de la section 4. Dans **(a)**, les astérisques indiquent des tests de Wilcoxon significatifs (* : p<0.05; ** : p<0.01; *** : p<0.001). Dans **(b)**, la différence d'âge prédit correspond à l'âge prédit moins l'âge réel.

4.3.2.3.2. Jeu d'entraînement multicentrique

Dans l'expérience de prédiction d'âge avec un jeu d'entraînement multicentrique, la procédure d'harmonisation a permis de faire passer l'EAM de 4.48 à 3.91 années. La différence d'erreur était significative d'après le test de Wilcoxon (p=0.0033). Les distributions d'erreurs pour chaque site sont illustrées dans l'annexe 7.8.

4.3.2.4. Corrélation entre l'âge et le volume de matière grise

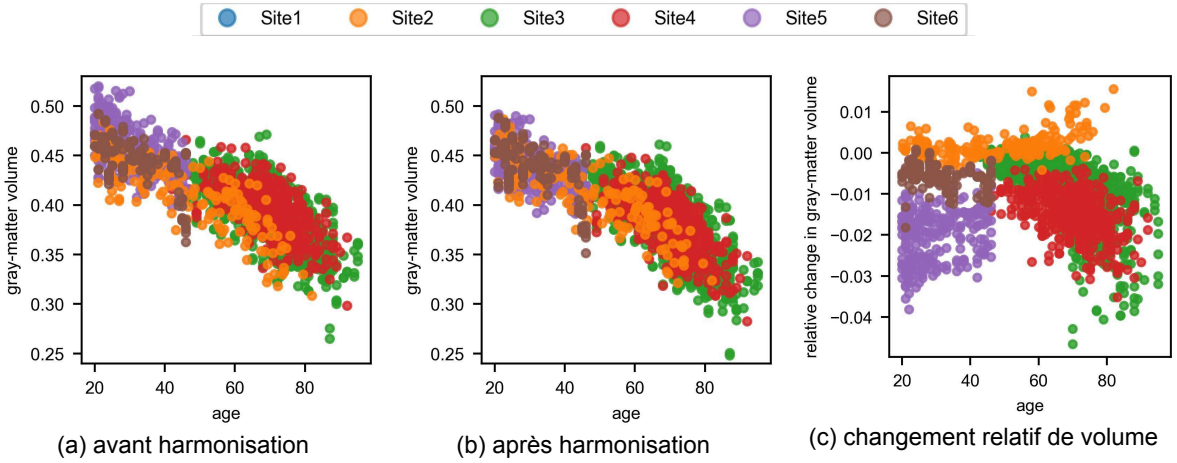
L'harmonisation a légèrement réduit la dispersion autour de la tendance centrale linéaire pour l'évolution des volumes de MG avec l'âge (Figures 38a et 38b). La corrélation linéaire a changé significativement : de -0.816 avant harmonisation à -0.821 après (p=0.0014 dans le test de Steiger). En observant l'évolution des estimations de volume (Figure 38c), on remarque qu'excepté pour Site2, les volumes étaient plus faibles après harmonisation. Pour Site3 et Site4, cet affaiblissement était d'autant plus grand avec l'âge.

4.3.2.5. Scores radiologiques

La consistance des cotations radiologiques entre les images avant et après harmonisation était excellente pour l'ACG (Figure 39a), l'ATM (Figure 39b), l'EPS-BG (Figure 39c), l'EPS-CS (Figure 39d) et l'index d'Evans (Figure 39e).

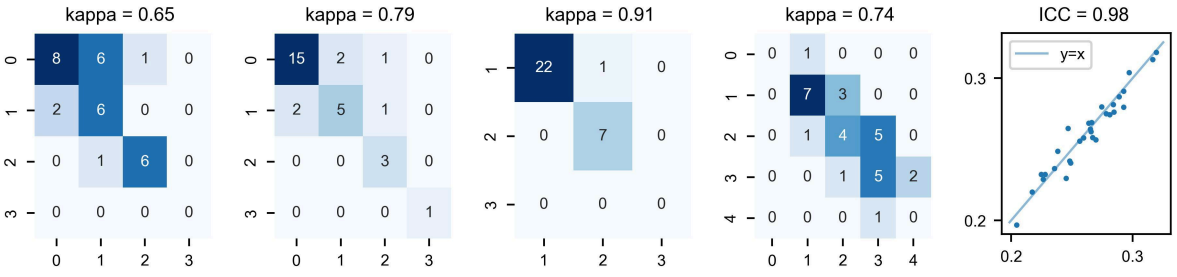
4.3.3. Harmonisation sur des sujets voyageurs

Le Tableau 6 reporte les SSIMs obtenus avec le jeu de données de sujets voyageurs. Notre modèle 3D a produit une augmentation significative du SSIM en harmonisant de Site3 vers Site4 ($p<0.001$) et une réduction significative en harmonisant de Site4 vers Site3 ($p<0.001$). Dans les deux cas, le CycleGAN 2D a réduit significativement les SSIMs ($p<0.001$). En outre, les SSIMs obtenus avec notre modèle 3D étaient significativement plus élevés dans les deux directions d'harmonisation ($p<0.001$).



(a) avant harmonisation (b) après harmonisation (c) changement relatif de volume

Figure 38 : Volume de matière grise en fonction de l'âge dans l'expérience multisite de la section 4. Les volumes sont divisés par le volumes intracrânien total. Dans **(c)**, le changement relatif est calculé comme le volume relatif après harmonisation moins celui avant harmonisation.



(a) atrophie corticale globale (b) atrophie temporale médiale (c) nombre d'espaces périvasculaires dilatés dans le ganglion de la base (d) nombre d'espaces périvasculaires dilatés dans le centre semi-ovale (e) index d'Evans

Figure 39 : Scores radiologiques dans l'expérience multisite de la section 4. Les résultats sont représentés dans un tableau de contingence pour les variables ordinales (la profondeur de couleur est proportionnelle à l'effectif dans chaque case). Les axes des Y et des X correspondent respectivement aux valeurs des images originales et harmonisées. ICC : corrélation intraclasse.

Tableau 6 : SSIMs dans le jeu de données de sujets voyageurs de la section 4.

	sans harmonisation	Site3 → Site4		Site4 → Site3	
		CycleGAN 2D	CycleGAN 3D	CycleGAN 2D	CycleGAN3D
SSIM [1]	0.9523 ± 0.0131	0.9407 ± 0.0111	0.9533 ± 0.0126	0.9454 ± 0.0105	0.9499 ± 0.0131

[1] Le SSIM est exprimé comme moyenne ± écart-type.

4.4. Discussion

Dans cette étude, nous avons développé un modèle 3D de transfert de domaine pour l'harmonisation inter-sites d'images cérébrales T1w. Étant donné les nombreuses préoccupations autour des exigences techniques des modèles d'apprentissage profond et de leur fiabilité avec diverses données IRM (Lundervold et Lundervold 2019), nous avons travaillé sur un usage efficient des ressources de calcul et une robustesse à différents jeux de données IRM. Basé sur un ensemble d'images IRM acquises avec différentes machines, le modèle a réduit des variabilités inter-sites de distributions d'intensités, de volumétrie cérébrale, de métriques de qualité et de radiomiques. L'harmonisation a également permis une prédiction d'âge significativement plus précise (que ce soit avec des jeux d'entraînements petits et unicentriques ou de grande taille et multicentriques) et atténué les effets de site sur le changement de volume de MG avec l'âge. Les scores radiologiques ont également confirmé la consistance des informations biologiques au niveau individuel avant et après harmonisation. De plus, une validation avec un jeu de données de sujets voyageurs a indiqué une supériorité par rapport à un modèle 2D reconnu.

Contrairement aux modèles précédemment développés qui traitent les images par partie (coupes ou patches 3D), celui présentement proposé a été conçu pour opérer sur des images cérébrales 3D T1w. Nous avons compensé les coûts supplémentaires en calcul en utilisant une architecture U-net pour les générateurs, moins gourmande que les architectures Resnet qui sont communes dans les approches CycleGAN (Chen et al. 2021; Gao et al. 2019; Nguyen et al. 2018; Palladino et al. 2020). Nous avons opté pour des convolutions transposées car l'alternative des *convolutions de redimensionnement* (Odena et al. 2016; Palladino et al. 2020) perdait plus d'informations anatomiques des images d'entrée (données non montrées) et produisaient des CNN plus lourds. Dewey et al. (2019) ont également jugé que les convolutions transposées étaient plus adaptées à l'harmonisation. Néanmoins, contrairement à celui de Dewey et al., l'architecture de nos générateurs est uniquement composée de convolutions avec des pas supérieurs à 1. Ce choix nous a permis de réduire les coûts calculatoires mais pourrait être une cause des quelques variabilités inter-sites restantes dans les IQMs après harmonisation (sections 4.3.1.2 et 4.3.2.2) étant donné que les convolutions transposées sont sujettes aux artefacts (Odena et al. 2016).

La structure CycleGAN de notre modèle nous a permis de traiter le problème des différences biologiques entre les populations des sites avec une stratégie d'échantillonnage biaisé mise en place pour éviter la suppression d'effets d'âge dans l'expérience multisite (section 4.2.4.1). La non-application de cette stratégie a mené à des pertes significatives de variabilités liées à l'âge (données non montrées) ce qui est cohérent avec des précédents résultats sur des tumeurs cérébrales (Cohen et al. 2018). Comparé à la suppression des images IRM ne rentrant pas dans une tranche d'âge spécifique (Dinsdale et al. 2021), cette méthode d'échantillonnage permet de mieux conserver la diversité des données d'entraînement. D'autres chercheurs ont récemment proposé des méthodes pour éviter la sur-correction causée par des déséquilibres biologiques durant l'entraînement de l'harmonisation : absence de l'information du site (Liu et al. 2021) ou intégration d'un module de conservation d'informations liées à des covariables (Cackowski et al. 2023).

Les résultats de nos expériences sites-appariés (section 4.3.1) confirment globalement l'efficacité du modèle avec divers jeux de données *source* et *cible* avec des différences de taille d'échantillon, de distribution d'âges, de scanner et de paramètres d'acquisition (section 4.2.1.1). Les données de Site5 et Site6 venaient de différentes études (i.e. NKI-RS et NMorphCH) et différaient significativement. Les données de Site1 et Site2 venaient de la même base de données (IXI) et étaient plus homogènes mais nos résultats ont quand même suggéré la présence de variabilités inter-sites claires, probablement dues à la différence d'intensité de champ (i.e. 1.5 vs 3T). Les différences entre Site3 et Site4 - deux jeux de données de l'étude OASIS-3 - étaient moins importantes mais notre modèle a tout de même réussi à en corriger de manière significative. C'est un point important, étant donné certaines harmonisations qui ont dégradé des informations biologiques dans des jeux de données avec de faibles effets de site (Richter et al. 2022).

L'expérience multisite nous a permis d'évaluer le modèle dans une situation plus commune où des données venant de plus de deux sites doivent être harmonisées vers un espace commun. Bien que des différences inter-sites étaient toujours présentes après harmonisation (principalement chez les 60-70 ans), notre analyse des distributions d'intensité (section 4.3.2.1) et de caractéristiques IRM (section 4.3.2.2) indiquent que les jeux de données ont bien été uniformisés. L'harmonisation a amélioré de manière significative les performances du modèle de prédiction d'âge entraîné sur les images de Site1 pour tous les sites, excepté Site5 (section 4.3.2.3.1). En considérant les différences d'âge prédit et la DME, la RME semblait plus importante après l'harmonisation des images de Site5. Nous avons alors fait l'hypothèse que les variabilités inter-sites ont mené à une sous-estimation de l'âge dans ce domaine et que cette sous-estimation a été fortuitement compensée par la RME qui tend à surestimer l'âge des jeunes participants. Nous avons validé cette hypothèse en incluant des participants plus vieux (annexe 7.9).

Ainsi, les résultats de prédiction d'âge mettent en avant l'intérêt de notre méthode pour l'adaptation de domaine, i.e. quand un modèle prédictif est entraîné sur des données d'un site et appliqué à d'autres (sections 4.3.1.3 et 4.3.2.3.1). De manière similaire, Bashyam et al. (2022) ont obtenu des améliorations significatives de prédiction d'âge avec leur modèle d'harmonisation. Nous sommes cependant allés plus loin et avons montré la valeur de notre modèle avec un jeu d'entraînement multicentrique de grande taille (section 4.3.2.3.2), ce qui n'était pas évident car la quantité et la diversité des données d'entraînement auraient pu favoriser la distinction entre les effets de site et d'âge et ainsi rendre l'harmonisation inutile, voir contre-productive. Robinson et al. (2020) ont réalisé une expérience similaire mais avec uniquement deux sites et leurs résultats ne rendent pas compte de l'importance des améliorations obtenues.

Les analyses sur la corrélation entre le volume de MG et l'âge (section 4.3.2.4) montrent l'effet du modèle sur un motif spécifique de vieillissement : la relation était significativement plus linéaire après harmonisation. Fortin et al. (2018) ont réalisé une expérience similaire avec des mesures d'épaisseurs corticales mais leur modèle d'harmonisation n'était pas appliqué au niveau de l'image. De plus, la corrélation initiale était plus faible (-0.70) que celle de l'étude présente (-0.82) et était donc plus simple à renforcer.

Les scores radiologiques ont montré que notre modèle est capable de conserver des informations radiologiques précises relatives à l'atrophie cérébrale, aux espaces périvasculaires et aux tailles des ventricules (sections 4.3.1.4 et 4.3.2.5). Dans des études précédentes de translation d'images médicales, le réalisme d'images générées a été évalué radiologiquement (Armanious et al. 2020; Welander et al. 2018). Nous avons plutôt choisi d'investiguer la conservation de caractéristiques individuelles avec l'harmonisation, étant

donné que les approches CycleGAN peuvent potentiellement produire des pertes d'informations des images d'origine (Cohen et al. 2018).

La validation sur le jeu de données de sujets voyageurs (section 4.3.3) montre que notre CycleGAN 3D est plus efficace que le CycleGAN 2D classique dans l'harmonisation des images de Site3 et Site4. L'harmonisation de Site3 vers Site4 a permis d'obtenir une augmentation significative du SSIM. Inversement, le SSIM a été significativement réduit avec l'harmonisation de Site4 à Site3. La faible variabilité inter-sites entre ces deux ensembles - d'après les différents résultats de l'expérience sites-appariés correspondante (section 4.3.1) - peut partiellement expliquer ce dernier résultat négatif. Par ailleurs, ce résultat va à l'encontre des résultats positifs obtenus en harmonisant ces deux sites, notamment ceux de la prédiction d'âge du modèle entraîné sur les images de Site3 et appliqué sur les images de Site4 (section 4.3.1.3).

Notre étude a quelques limites. Premièrement, l'approche CycleGAN requiert un entraînement indépendant pour chaque site. Néanmoins, la grande diversité des jeux de données et des évaluations utilisés dans cette étude comparé à la littérature montre l'intérêt de notre modèle pour des jeux de données comprenant un nombre relativement important d'images IRM. Gebre et al. (2023) ont également montré l'intérêt de CycleGAN par rapport à des approches ComBat, des GANs conditionnels et des méthodes de transfert de style pour l'harmonisation d'ensembles de données transversales de grande taille. Deuxièmement, nous avons choisi de comparer notre approche avec un CycleGAN 2D car ce dernier est un modèle reconnu pour lequel des détails d'implémentation clairs sont fournis. Cependant, l'absence de comparaison avec une approche plus récente est une limite. Nous avons fait ce choix car il aurait été très difficile de réaliser une comparaison équitable à cause de plusieurs facteurs : l'absence de code accessible ou facilement utilisable (également mis en avant par Hu et al. (2023)), l'hétérogénéité dans les données utilisées et dans les prétraitements (section 3.3). Troisièmement, beaucoup de caractéristiques IRM que nous avons utilisées pour analyser les effets de site sont dépendantes de la méthode de segmentation (ici FSL-FAST) et nos résultats auraient pu être différents si une autre avait été appliquée. Toutefois, l'utilisation de nombreuses mesures a fourni des informations pertinentes et limité l'impact de l'outil de segmentation. Quatrièmement, malgré la nature générique de notre méthode, nous avons uniquement étudié l'harmonisation d'images cérébrales T1w. Bien que les séquences T1w soient utilisées dans beaucoup de travaux de recherche et en pratique clinique (e.g. pour la démence et la sclérose en plaques), l'utilisation d'approches multimodales peut améliorer l'harmonisation (Dewey et al. 2019). Enfin, nous avons uniquement appliqué notre modèle sur des participants sains. Nous avons maintenant l'intention de le tester sur des données de personnes atteintes de troubles neurologiques[3]. Nous prévoyons également d'étendre notre cadre CycleGAN pour combiner sa capacité d'harmonisation avec la facilité des méthodes qui peuvent être appliquées à n'importe quelle image après l'entraînement[3].

4.5. Conclusion

Dans cet article, nous proposons un modèle non supervisé pour l'harmonisation inter-sites d'images IRM 3D T1w du cerveau. Cette approche d'apprentissage profond 3D optimisée traite des images du cerveau entier et est robuste face à divers jeux de données IRM. Diverses expériences menées sur différentes cohortes à différentes échelles témoignent de

[3] Ces projections ont été mises en œuvre dans l'étude présentée en section 5.

la capacité du modèle à éliminer différents types de variabilités inter-sites, à conserver les informations radiologiques et à renforcer des motifs biologiques. Malgré la présence de différences biologiques majeures entre les sites, notre choix d'une stratégie d'entraînement appropriée a contribué à la réussite de l'harmonisation multisite. Notre validation approfondie des résultats d'harmonisation est prometteuse pour diverses applications futures dans le cadre d'études multicentriques.

5. IGUANe : un CycleGAN 3D généralisable pour une harmonisation multicentrique d'images IRM cérébrales structurelles

Résumé

Nous avons vu précédemment que des modèles d'apprentissage profond pour de la translation d'images ont été adaptés pour de l'harmonisation inter-sites en IRM. Dans ce chapitre, nous proposons IGUANe (Image Generation with Unified Adversarial Networks), un modèle 3D original qui combine les forces de la translation de domaine avec la facilité d'application des méthodes de transfert de style pour une harmonisation inter-sites d'images IRM cérébrales. IGUANe étend CycleGAN en intégrant un nombre arbitraire de domaines pendant l'entraînement grâce à une stratégie *plusieurs-à-un*. À l'inférence, le modèle peut être appliqué à n'importe quelle image même si le site d'origine n'est pas connu et constitue donc un générateur universel pour l'harmonisation. Après un entraînement sur un jeu de données comprenant des images T1w de 11 scanners différents, des expériences sur des images venant d'autres études ont évalué (i) la transformation d'images IRM d'un domaine à un autre avec des sujets voyageurs, (ii) la conservation des différences entre les images IRM d'un même domaine, (iii) l'évolution de motifs volumétriques liés à l'âge et à la maladie d'Alzheimer et (iv) l'évolution de performances de prédiction d'âge et de classification de patients. Des comparaisons avec d'autres méthodes d'harmonisation et de normalisation suggèrent qu'IGUANe harmonise plus efficacement les images IRM et est plus robuste à diverses applications subséquentes. Le modèle entraîné pour cette étude peut être utilisé très facilement sur toute image cérébrale T1w. D'autres chercheurs pourraient aussi réentraîner un modèle pour une harmonisation de différents types d'images.

Mots clés : IRM cérébrale ; harmonisation ; volumétrie cérébrale ; âge cérébral ; maladie d'Alzheimer

5.1. Introduction

Nous avons précédemment introduit l'IRM et la problématique de la variabilité inter-sites (section 1). Nous avons également fait une revue des différents types de méthodes pour l'harmonisation rétrospective (section 2). Dans ce chapitre, nous nous intéressons aux modèles d'apprentissage profond non-supervisés pour la translation d'images et l'harmonisation d'images IRM cérébrales.

Parmi ces approches, nous distinguons les méthodes de translation de domaine et de transfert de style. La plupart des méthodes les plus récentes sont basées sur du transfert de style (Cackowski et al. 2023; Liu et al. 2023; Zuo et al. 2021b, 2022). Un argument en faveur de ces méthodes est leur fiabilité par rapport aux méthodes de translation de domaine qui sont sujettes à la sur-correction (Cohen et al. 2018). La possibilité d'application à des images IRM provenant de sites non vus est un autre avantage présumé. Cependant, les quelques expériences qui l'ont attestée étaient limitées à une évaluation visuelle et à des mesures de similarité d'images avec des jeux de données de sujets voyageurs (Cackowski et al. 2023; Liu et al. 2023). D'autre part, les modèles de translation de domaine - souvent basés sur une architecture CycleGAN (Zhu et al. 2017) - cherchent à apprendre des correspondances de plus haut niveau entre des ensembles d'images et ont montré leur

efficacité pour l'harmonisation inter-sites (Chen et al. 2021; Gebre et al. 2023; Palladino et al. 2020; section 4). De plus, des recherches précédentes ont suggéré des améliorations potentielles pour rendre les modèles de translation de domaine généralisables aux sites non vus (Gao et al. 2019) et pour prévenir la sur-correction (section 4).

L'évaluation des modèles est également un aspect clé de l'harmonisation. Nous avons déjà mentionné l'utilisation de métriques de similarité sur des jeux de données de sujets voyageurs (section 3.2.2.1). De tels datasets fournissent des vérités terrains et sont très utiles pour l'harmonisation. Toutefois, ces validations ne sont applicables que sur quelques petits jeux de données et ne sont pas forcément en adéquation avec la qualité d'harmonisation (sections 3.2.2.1 et 3.3). Elles sont alors souvent complétées par des évaluations de similarités entre les sites avant et après harmonisation (section 3.2.3). Mais ces dernières ne prennent pas en compte les informations biologiques et sont seulement applicables à des images IRM qui peuvent être regroupées dans des domaines (e.g. des sites). Pour surmonter ces limites, l'étude des performances de modèles prédictifs (section 3.2.5) et de l'évolution de motifs biologiques spécifiques avec l'harmonisation (section 3.2.4) peut être réalisée.

Dans ce chapitre, nous proposons IGUANe (Image Generation with Unified Adversarial Networks), un nouveau modèle génératif pour l'harmonisation inter-sites d'images IRM structurelles du cerveau qui combine la puissance de la translation de domaine avec la praticité du transfert de style. IGUANe est basé sur de l'entraînement antagoniste avec une contrainte de consistance du cycle. Après avoir été entraîné sur un jeu de données multicentrique de grande taille, il permet l'harmonisation de toute image IRM, peu importe le site d'acquisition. Pour valider la robustesse de la méthode proposée, nous l'avons testée en utilisant des jeux de généralisation qui contiennent diverses images IRM de différents constructeurs avec des caractéristiques d'acquisition variables. Nous avons utilisé des modèles prédictifs et nous avons analysé des motifs biologiques dans les images pour quantifier l'évolution de l'information biologique avec l'harmonisation, en comparant avec deux méthodes récentes de transfert de style et deux méthodes de normalisation bien établies. Les résultats montrent que, comparé à ces approches, IGUANe est robuste face à une variété d'analyses IRM.

5.2. Matériels et méthodes

5.2.1. Jeux de données IRM

Nous avons utilisé des images T1w du cerveau venant de bases de données publiques pour cette étude. Nous avons uniquement inclus des participants entre 18 et 80 ans. Les métadonnées permettant l'identification des images sont disponibles dans notre répertoire en ligne (section 5.2.5). Quelques images IRM ont été exclues à cause de la taille du cerveau (section 5.2.2.1). Les caractéristiques de chaque jeu de données sont décrites ci-dessous.

Harmonization dataset. Dans cette collection, nous avons inclus des images IRM provenant de huit études : SALD (Wei et al. 2018), IXI[4], OASIS-3 (LaMontagne et al. 2019), NKI-RS (Nooner et al. 2012), NMorphCH[5], AIBL (Ellis et al. 2009), HCP Young Adult[6] et

[4] https://brain-development.org/ixi-dataset/ accédé le 15/01/2022
[5] http://otto.fsm.northwestern.edu/documentation accédé le 22/02/2019
[6] https://www.humanconnectome.org/study/hcp-young-adult accédé le 22/02/2019

ICBM[7]. Comme spécifié dans les protocoles des études, tous les participants étaient des sujets sains. Onze machines ont acquis les images et représentaient les domaines pour notre modèle d'harmonisation (section 5.2.2.6). Chaque participant était présent dans un seul domaine. Des informations démographiques et sur les scanners sont données dans le Tableau 7.

Tableau 7 : Caractéristiques des scanners et des participants dans *Harmonization dataset*.

Étude	Modèle du scanner	Intensité du champ, Tesla	Nombre d'images IRM / de participants	Âge, années [1]	Hommes, %
SALD	Siemens Magnetom TrioTim	3	494/494	45.18 ± 17.44	38
IXI	Philips Intera	1.5	305/305	47.31 ± 16.52	48
IXI	Philips Intera	3	176/176	50.20 ± 15.46	43
OASIS-3	Siemens Magnetom TrioTim	3	857/350	67.03 ± 8.27	33
OASIS-3	Siemens BioGraph mMR PET-MR	3	412/311	68.10 ± 7.51	46
NKI-RS	Siemens Magnetom TrioTim	3	249/247	29.97 ± 8.20	40
NMorphCH	Siemens Magnetom TrioTim	3	141/44	31.37 ± 8.42	53
AIBL	Siemens Magnetom TrioTim	3	489/280	71.93 ± 4.90	46
HCP	Siemens Connectome Skyra	3	402/402	28.90 ± 3.71	39
ICBM	Siemens Sonata	1.5	677/135	43.97 ± 15.21	48
ICBM	Philips ACS III	1.5	145/145	25.10 ± 4.93	56

[1] L'âge est exprimé comme moyenne ± écart-type.

SPRBS_TS. Le jeu de données de sujets voyageurs de SRPBS (Tanaka et al. 2021) inclut des images IRM cérébrales de neuf hommes sains de 24 à 32 ans. Dans notre étude, nous

[7] https://ida.loni.usc.edu accédé le 15/04/2023

avons utilisé un total de 97 images acquises avec 11 machines (constructeurs GE, Siemens et Philips) pour évaluer l'harmonisation de manière supervisée (section 5.2.4.1).

Generalization dataset. Nous avons inclus dans ce jeu de données des images IRM de cinq autres études : ADNI[8], MCIC (Gollub et al. 2013), PPMI (Marek et al. 2018), COBRE (Aine et al. 2017) et ABIDE[9]. Comme spécifié dans les protocoles des études, tous les participants étaient des sujets sains. Des informations démographiques et sur les scanners sont données dans le Tableau 8. Nous avons utilisé ces images pour évaluer l'harmonisation sur des sujets sains (sections 5.2.4.2 et 5.2.4.4). Il doit être noté qu'aucun sujet ni aucun scanner n'était présent lors de l'entraînement du modèle d'harmonisation.

Tableau 8 : Caractéristiques des scanners et des participants dans *Generalization dataset*.

Étude	Constructeur [1]	Intensité du champ, Tesla [1]	Nombre d'images IRM / de participants	Âge, années [2]	Hommes, %
ADNI	GE (115); Philips (109); Siemens (104)	3 (199); 1.5 (129)	328/216	73.62 ± 4.37	45
MCIC	Siemens	1.5 (192); 3 (52)	244/89	34.49 ± 11.93	67
PPMI	Siemens (202); Philips (39)	3 (215); 1.5 (24); unknown (2)	241/141	60.49 ± 10.65	64
COBRE	Siemens	3	227/91	38.57 ± 11.55	73
ABIDE	Siemens (105); Philips (34)	3	139/139	27.40 ± 6.83	88

[1] Le nombre d'images IRM est indiqué entre parenthèses si il y a plusieurs options.
[2] L'âge est exprimé comme moyenne ± écart-type.

AD dataset. Nous avons utilisé des données de ADNI et AIBL pour étudier l'évolution de motifs liés à la maladie d'Alzheimer avec l'harmonisation. Nous avons sélectionné des participants qui ont été diagnostiqués normaux cognitivement (CN) ou Alzheimer (AD). Pour mener une classification CN/AD (section 5.2.4.4), nous avons divisé la collection entre trois ensembles (≠ participants) : **AD_train** pour l'entraînement, **AD_test** pour le test et **AD_gen** comme jeu de généralisation d'images IRM du constructeur GE (aucune image GE n'était dans *AD_train*, *AD_test* ni *Harmonization dataset*). Il doit être noté qu'aucun sujet ni aucun scanner n'était présent lors de l'entraînement du modèle d'harmonisation.

Des informations démographiques et sur les scanners sont données dans le Tableau 9.

[8] https://adni.loni.usc.edu/about/ accédé le 15/04/2023
[9] https://fcon_1000.projects.nitrc.org/indi/abide/ accédé le 15/04/2023

Tableau 9 : Caractéristiques des scanners et des participants dans *AD dataset*.

Jeu de données	Constructeur [1]	Intensité du champ, Tesla [1]	Nombre d'images IRM / de participants	CN/AD, %	Âge, années [2]	Hommes, %
AD_train	Siemens (1286); Philips (489)	3 (1132); 1.5 (643)	1775/546	50/50	72.23 ± 5.54	49
AD_test	Siemens (518); Philips (169)	3 (417); 1.5 (270)	687/237	50/50	71.79 ± 5.90	49
AD_gen	GE	1.5 (558); 3 (546)	1104/261	47/53	72.94 ± 4.74	49

[1] Le nombre d'images IRM est indiqué entre parenthèses si il y a plusieurs options.
[2] L'âge est exprimé comme moyenne ± écart-type.

5.2.2. Modèle IGUANe

Le code pour l'entraînement et l'inférence de IGUANe est accessible en ligne avec les poids du modèle entraîné pour cette étude (section 5.2.5). Des détails d'implémentation sont également donnés dans l'annexe 7.10.

5.2.2.1. Prétraitements IRM

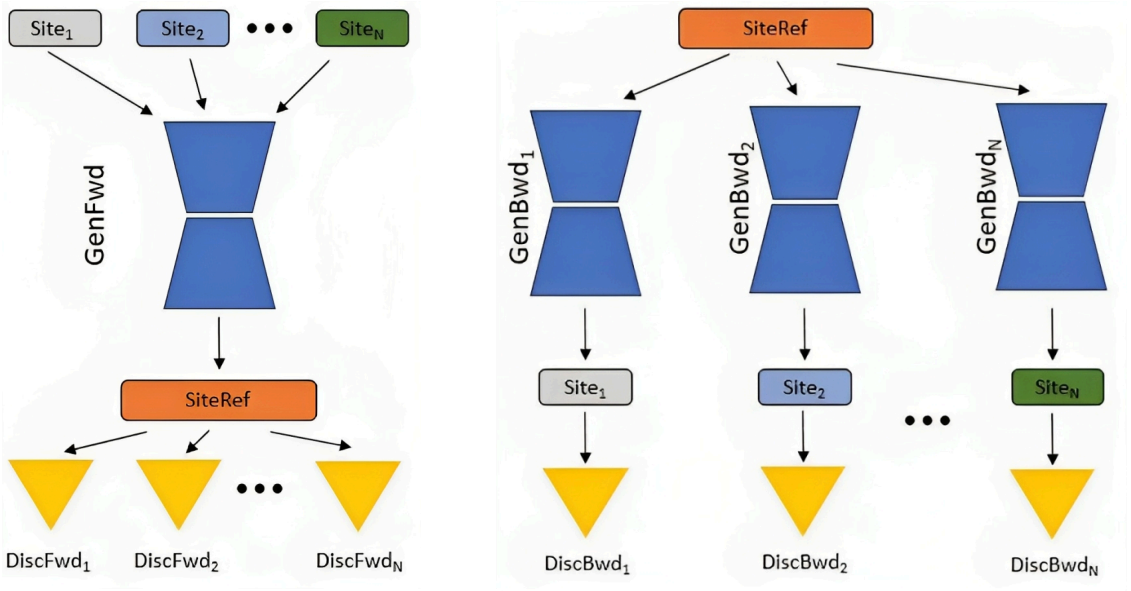
En vue d'éliminer les variabilités techniques les plus triviales dans les données et pour faciliter l'harmonisation, nous avons prétraité les images IRM en suivant les étapes suivantes : (i) extraction de la boîte crânienne avec HD-BET (Isensee et al. 2019), (ii) correction d'inhomogénéités de champ avec N4ITK (Tustison et al. 2010), (iii) recalage linéaire vers espace du MNI 1 mm^3 avec FSL-FLIRT (Jenkinson et al. 2002) (six degrés de liberté), (iv) rognage vers un espace de 160 x 192 x 160 voxels (1.8% d'images exclues dans *Harmonization dataset*) et (v) division des intensités par l'intensité médiane dans le cerveau. Cette dernière étape standardise l'intensité médiane dans le cerveau tout en conservant l'arrière-plan à 0. Nous avons trouvé cela plus robuste face aux valeurs aberrantes que la normalisation min-max (section 2.3.1). Pour IGUANe et pour les évaluations reposant sur des modèles d'apprentissage profond (section 5.2.4.4), les intensités ont été mises à l'échelle et décalées avec des constantes de telle sorte que l'arrière-plan et la médiane étaient respectivement à -1 et 0 ; sinon, elles ont été mises à 0 et 500.

5.2.2.2. Générateur universel

L'architecture de IGUANe est une extension de celle de CycleGAN (Zhu et al. 2017) et a été inspirée par la stratégie *plusieurs-à-un* de Gao et al. (2019). L'objectif est d'entraîner un générateur *GenFwd* à translater des images de *N* sites source ($Site_1$, $Site_2$, … et $Site_N$) vers un site de référence *SiteRef*. Pour permettre des contraintes de consistance du cycle comme dans CycleGAN, *N* générateurs inversés ($GenBwd_1$, $GenBwd_2$, … et $GenBwd_N$) sont

établis pour translater des images de *SiteRef* vers chaque site source. Pour l'entraînement antagoniste, *N* discriminateurs (*DiscFwd*$_1$, *DiscFwd*$_2$, … et *DiscFwd*$_N$) apprennent à distinguer les images réelles de *SiteRef* et les images harmonisées de chaque site source avec *GenFwd*. De la même manière, *N* discriminateurs inversés (*DiscBwd*$_1$, *DiscBwd*$_2$, … et *DiscBwd*$_N$) apprennent à distinguer les images réelles de chaque site source des images harmonisées avec les générateurs inversés.

Les modules de l'architecture sont illustrés dans la Figure 39. À l'inférence, *GenFwd* est utlisé pour harmoniser des images IRM. L'hypothèse est qu'un entraînement avec des données suffisamment diverses permet d'harmoniser n'importe quelle image, y compris quand le site d'acquisition est inconnu.



(a) harmonisation vers *SiteRef*

(b) harmonisation vers les sites source

Figure 39 : Représentation des modules dans l'architecture IGUANe.

5.2.2.3. Fonctions de coût

L'erreur quadratique moyenne est utilisée comme fonction de coût antagoniste (Mao et al. 2017). Pour générer des images de *SiteRef* réalistes, *GenFwd* est entraîné successivement de manière antagoniste avec *DiscFwd*$_1$, *DiscFwd*$_2$, … et *DiscFwd*$_N$. Les *N* générateurs inversés sont entraînés de la même manière avec chaque discriminateur avec *DiscBwd*$_1$, *DiscBwd*$_2$, … et *DiscBwd*$_N$.

Pour préserver les informations de l'image originale et pour régulariser l'entraînement des générateurs, des fonctions de coût de consistance du cycle et d'identité sont utilisées (Zhu et al. 2017). Les consistances du cycle sont calculées en faisant collaborer *GenFwd* avec chaque générateur inversé avec translations et translations inversées dans les deux directions pour chaque site.

5.2.2.4. Architectures des réseaux

Les architectures des générateurs suivent celles décrites par Dewey et al. (2019), qui utilise des *connexions sautées* pour favoriser la conservation de détails anatomiques. Une différence notable est que des convolutions 3D sont utilisées dans IGUANe à la place des 2D afin de traiter des images cérébrales complètes. De plus, pour mieux préserver le

contenu des images originales, la tâche des générateurs a été modifiée vers un apprentissage résiduel, ce qui signifie que les sorties des générateurs sont additionnées voxel à voxel aux entrées pour obtenir l'image harmonisée finale (de Bel et al. 2021). La dernière fonction d'activation avant l'addition est *tanh* pour permettre des résidus négatifs ou positifs. À l'inférence, les voxels négatifs sont mis à la valeur de l'arrière-plan.

Les discriminateurs sont des classifieurs patchGAN (Isola et al. 2017) avec des convolutions 3D pour un champ de réception de 54^3 voxels.

Étant donné que le modèle a été conçu pour traiter des images IRM sans boîte crânienne (section 5.2.2.1), une partie importante des images est de l'arrière-plan. Par défaut, l'arrière-plan est associé avec l'intensité minimale dans les images. Dans IGUANe, l'arrière-plan est mis à 0 - ce qui correspond à l'intensité cérébrale médiane (section 5.2.2.1) - avant que les images ne soient données aux générateurs et aux discriminateurs. L'idée est d'affecter une valeur plus *neutre* à une partie de l'image qui a une taille importante mais qui ne devrait pas être prise en compte par l'harmonisation. De plus, le masque du cerveau original est appliqué après chaque translation d'images pour faire en sorte que les générateurs se focalisent sur les intensités cérébrales durant l'entraînement (section 4.2.3.1).

Pour les générateurs et les discriminateurs, des normalisations d'instance (Ulyanov et al. 2017) avec des déviations absolues moyennes au lieu d'écarts-types (Wu et al. 2019) sont utilisées.

5.2.2.5. Procédure d'entraînement

À chaque étape d'entraînement, les sous-étapes suivantes sont suivies pour chaque index de site source *i* (séquence dans un ordre aléatoire).

1. Entraînement des discriminateurs :
    a. 4 images sont sélectionnées dans *SiteRef* et dans $Site_i$.
    b. 2 images de *SiteRef* et 2 de $Site_i$ translatées vers *SiteRef* avec *GenFwd* sont utilisées pour mettre à jour $DiscFwd_i$.
    c. Les 2 autres images de *SiteRef* translatées vers $Site_i$ avec $GenBwd_i$ et les 2 autres de $Site_i$ sont utilisées pour mettre à jour $DiscBwd_i$.
2. Entraînement des générateurs :
    a. 1 image est sélectionnée de *SiteRef* et 1 de $Site_i$.
    b. *GenFwd* et $GenBwd_i$ sont mis à jour avec les deux images en calculant les fonctions de coût antagonistes (en utilisant respectivement $DiscFwd_i$ et $DiscBwd_i$), les fonctions de coût de consistance du cycle et d'identité.

5.2.2.6. Implémentation

Nous avons entraîné IGUANe en utilisant *Harmonization dataset*. Nous avons sélectionné SALD comme le site de référence de par son grand nombre d'images IRM et sa large gamme d'âges (19-80 ans). Les dix autres sites de *Harmonization dataset* étaient alors les sites source. Étant donné les différences de distribution d'âges entre SALD et chaque site source (Tableau 7), nous avons mis en place une stratégie d'échantillonnage pour chacun de telle sorte qu'ils eussent une distribution d'âges similaire à celle de SALD (section 4.2.4.1). Une distribution de probabilités basée sur l'âge des participants était ainsi effective pour échantillonner les images IRM pendant les sous-étapes d'entraînement associées à chaque site source (sous-étapes 1a et 2a dans la section 5.2.2.5). Les distributions de probabilités utilisées dans cette étude peuvent être visualisées dans l'annexe 7.5.3.

Pour encourager davantage la préservation des informations biologiques, nous avons implémenté une procédure de validation qui évalue le modèle toutes les 5 époques et enregistre la meilleure version. À cette fin, deux modèles d'apprentissage profond pré-entraînés pour la prédiction de l'âge et du sexe à l'aide d'images IRM de *SiteRef* ont été appliqués à un sous-ensemble d'images IRM de chaque site source harmonisé vers *SiteRef*. L'architecture du modèle décrite par Cole et al. (2017) a été utilisée.

Dans cette étude, l'entraînement d'IGUANe a consisté en l'optimisation parallèle de 11 générateurs et 20 discriminateurs, ce qui implique un coût important en ressources de calcul. Pour sauvegarder de la mémoire GPU et accélérer les calculs, une politique de précision mixte a été suivie (Micikevicius et al. 2018) pour permettre l'entraînement sur un GPU NVIDIA Quadro RTX 6000 (mémoire 24 Gb).

5.2.3. Comparaison avec d'autres approches

5.2.3.1. Techniques de normalisation

L'égalisation d'histogrammes (HM) a été beaucoup utilisée dans des études IRM multicentriques (section 2.4.1). Dans cette étude, nous avons utilisé les images IRM SALD comme jeu d'entraînement utilisé pour déterminer l'échelle d'intensité standard vers laquelle les histogrammes d'intensité sont harmonisés.

Nous avons également implémenté la normalisation WhiteStripe (WS), qui comme HM, a été utilisée dans plusieurs études multicentriques pour la normalisation d'intensité (section 2.3.3).

Pour HM et WS, nous avons prétraité les images de la même manière que pour IGUANe (section 5.2.2.1).

5.2.3.2. Approches de transfert de style

Pour tester l'approche de Liu et al. (2023), nous avons utilisé le code fourni en ligne avec le modèle pré-entraîné[10]. Leur modèle (STGAN) est basé sur une image de référence de laquelle un code de style est extrait et utilisé comme cible pour l'harmonisation. Ici, nous avons utilisé l'image de référence fournie dans le dépôt en ligne pour nos expériences. Nous avons reproduit les prétraitements détaillés dans l'article de Liu et al. et avons ajouté une normalisation min-max (non mentionnée dans l'article mais effective dans le code en ligne).

Nous avons aussi testé l'approche de Zuo et al. (2021b, 2021a) (CALAMITI) avec le code et le modèle pré-entraîné accessible en ligne[11]. Nous avons utilisé l'image T1w 3D fournie comme référence pour l'harmonisation. Nous avons suivi les étapes de prétraitement données en ligne. Pour le recalage, nous avons utilisé FSL-FLIRT avec six degrés de liberté et un template 0.8 mm^3. Après WS (dernière étape de prétraitement de CALAMITI), nous avons mis à l'échelle et décalé les intensités IRM pour faire correspondre la moyenne et l'écart-type de la MB apparaissant normale (Shinohara et al. 2014) avec ceux de l'image de référence (détails dans l'annexe 7.11).

[10] https://github.com/USCLoBeS/style_transfer_harmonization accédé le 15/07/2023
[11] https://iacl.ece.jhu.edu/index.php?title=CALAMITI accédé le 27/07/2023

5.2.4. Expériences

5.2.4.1. Différences inter-sites et inter-sujets

Nous avons utilisé les sujets voyageurs de SRPBS_TS pour évaluer la capacité des modèles d'harmonisation à transformer des images d'un site vers leur équivalent dans un autre. Pour ce faire, nous avons calculé pour chaque sujet le SSIM (Wang et al. 2004) pour chaque paire d'images (correspondant à chaque paire de sites). Avant les calcus de SSIM, nous avons enregistré linéairement les images IRM avec FSL-FLIRT (six degrés de liberté) vers l'image IRM du sujet correspondant du site HKH. La gamme dynamique pour le SSIM a été définie comme le 99e percentile des intensités de voxel des deux images IRM.

Pour compléter l'expérience, nous avons suivi la stratégie de Liu et al. (2023) pour quantifier la conservation des différences inter-sujets avec l'harmonisation. Plus précisément, nous avons calculé au sein de chaque site SRPBS_TS la distance euclidienne pour chaque paire d'images (correspondant à chaque paire de sujets) avant et après l'harmonisation. Nous avons ensuite calculé pour chaque site le coefficient de corrélation de Pearson entre les distances avant et après harmonisation.

Pour ces évaluations, nous avons comparé IGUANe avec STGAN uniquement car les autres approches impliquent des valeurs d'arrière-plan variables, ce qui rend difficile la comparaison de mesures de similarité au niveau de l'image.

5.2.4.2. Corrélation entre l'âge et le volume de matière grise

De manière similaire à l'étude précédente (section 4.2.4.2.3), nous avons calculé les coefficients de corrélation de Pearson entre l'âge et le volume de MG avant et après harmonisation. Nous avons cette fois utilisé l'outil de segmentation SPM12 et pour chaque image de *Generalization dataset*, nous avons estimé le volume de MG que nous avons divisé par le volume intracrânien total. Nous avons aussi appliqué une régression linéaire des moindres carrés pour quantifier la perte de MG avec l'âge (l'âge était la variable indépendante).

Nous avons comparé IGUANe avec STGAN et CALAMITI pour cette évaluation mais pas avec HM et WS car ces deux méthodes produisent des arrière-plans négatifs qui altèrent les segmentations SPM.

5.2.4.3. Volumes hippocampiques : taille de l'effet cas/témoin

Une perte accélérée de volume hippocampique est un motif bien connu de la maladie d'Alzheimer (Schuff et al. 2009). Pour étudier ce motif, nous avons sélectionné aléatoirement 250 participants CN et 250 participants AD dans *AD dataset* (moyenne d'âge 71.34 et 71.90 dans les groupes CN et AD, respectivement). Ensuite, de manière similaire à Liu et al. (2023), nous avons segmenté les images IRM avec SynthSeg (Billot et al. 2023) et avons calculé le d de Cohen entre les volumes hippocampiques des participants CN et AD.

5.2.4.4. Tâches de prédiction

Nous avons évalué l'effet de l'harmonisation sur deux tâches de prédiction : la régression d'âge et la classification CN/AD. La régression d'âge consiste à entraîner un modèle à prédire l'âge d'un individu à partir de données d'IRM cérébrale et a été beaucoup expérimentée avec des modèles d'apprentissage profond (Cole et al. 2017; Gautherot et al.

2021; Jonsson et al. 2019). De telles approches ont aussi été appliquées à la classification dans la maladie d'Alzheimer (Basaia et al. 2019; Vieira et al. 2017).

Ici, nous avons implémenté des régresseurs d'âge et des classifieurs CN/AD basés sur des images IRM en suivant l'architecture de réseau proposée par Cole et al. (2017). Pour entraîner les modèles de prédiction d'âge, nous avons sélectionné aléatoirement 2178 images IRM de *Harmonization dataset* en conservant la proportion de chaque site. Nous avons ensuite appliqué le modèle sur les images de *Generalization dataset*. Le classifieur CN/AD a été entraîné avec les images de *AD_train* puis évalué sur les images de *AD_TEST* et *AD_GEN*. Des détails supplémentaires relatifs aux modèles de prédiction sont donnés dans l'annexe 7.12.

Nous n'avons pas inclus STGAN et CALAMITI dans ces expériences car les images IRM sont bien grandes (environ x3.4) et auraient donc impliqué des modèles bien plus lourds très difficiles à entraîner. Ainsi, un modèle de prédiction d'âge et un classifieur CN/AD ont été entraînés et évalués sur les images IRM prétraitées (prétraitements IGUANe) and sur les images IRM après les harmonisations IGUANe, HM et WS (4 modèles de prédiction d'âges et 4 classifieurs CN/AD).

5.2.5. Accès aux données et au code

Le code et les poids du modèle d'harmonisation, ainsi que les métadonnées relatives aux images IRM utilisées dans cette étude seront accessibles prochainement en ligne.

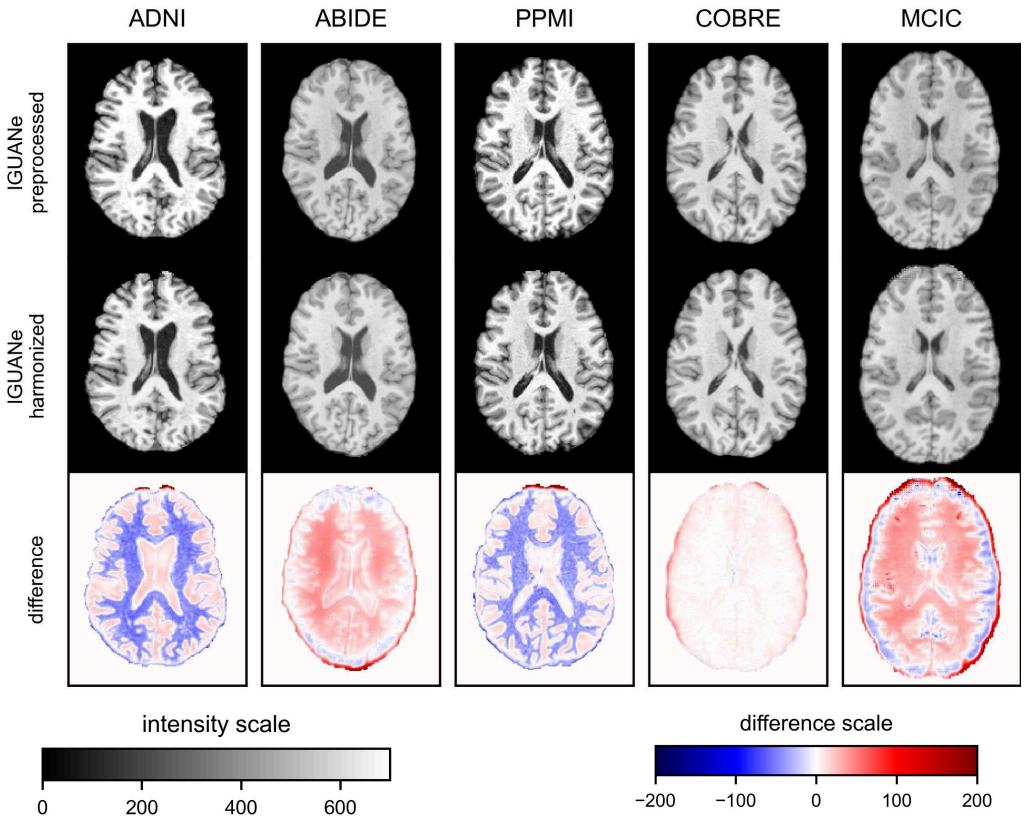


Figure 40 : Visualisation d'harmonisations IGUANe. Une image a été sélectionnée aléatoirement dans chaque étude de *Generalization dataset* et la coupe axiale centrale est montrée. Les cartes de différence correspondent à une soustraction voxel à voxel, i.e. les images harmonisées moins les originales.

5.3. Résultats

5.3.1. Évaluation visuelle de l'harmonisation

L'effet de l'harmonisation IGUANe est illustrée dans la Figure 40. Comme nous avons conçu le modèle pour une application à des images T1w de toute source d'acquisition et que nous voulions d'abord préserver les informations anatomiques, la visualisation des changements de contraste n'est pas évidente. Cependant, les cartes de différence montrent qu'en fonction de l'image d'entrée, IGUANe est capable d'appliquer différentes modifications des contrastes d'entrée. Par exemple, on peut voir que IGUANe a réduit le contraste MG/MB pour les images ADNI et PPMI, alors qu'il a été plutôt augmenté pour les images ABIDE et MCIC.

5.3.2. Évaluations sur des sujets voyageurs

Le Tableau 10 présente les SSIMs obtenus entre les images IRM d'un même sujet dans SRPBS_TS. Le SSIM n'a pas augmenté avec IGUANe alors qu'il était légèrement plus grand après STGAN.

Tableau 10 : SSIM intra-sujet dans SRPBS_TS.

	IGUANe		STGAN	
	prétraitées	harmonisées	prétraitées	harmonisées
SSIM [1]	0.916 ± 0.033	0.916 ± 0.031	0.940 ± 0.021	0.948 ± 0.022

[1] Le SSIM est exprimé comme moyenne ± écart-type.

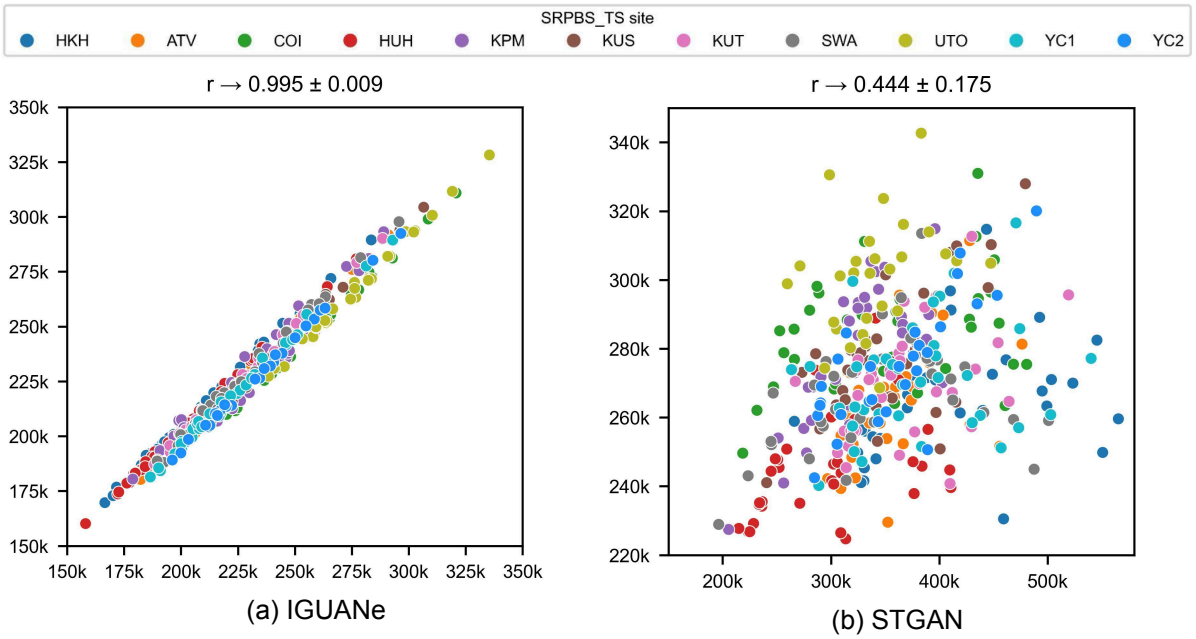


Figure 41 : Distances inter-sujets dans SRPBS_TS avant et après harmonisation. Les axes des X et des Y indiquent les distances Euclidiennes avant et après harmonisation, respectivement.

En revanche, IGUANe a mieux préservé les différences inter-sujets que STGAN (Figure 41).

5.3.3. Corrélation entre l'âge et volume de matière grise

Dans la Figure 42, on peut voir que la plus forte corrélation négative a été obtenue avec l'harmonisation IGUANe. IGUANe a également renforcé la pente de régression, indiquant que le motif de perte de MG a été renforcé par l'harmonisation. Une corrélation plus importante a également été obtenue avec STGAN, mais elle était moins significative que celle obtenue avec IGUANe. Contrairement à IGUANe, STGAN a produit une réduction de la pente de régression. Par ailleurs, CALAMITI a réduit la linéarité de la corrélation et a affaibli la pente de régression. On peut également observer que la corrélation linéaire obtenue après le prétraitement STGAN était plus forte que celles obtenues après les prétraitements IGUANe et CALAMITI.

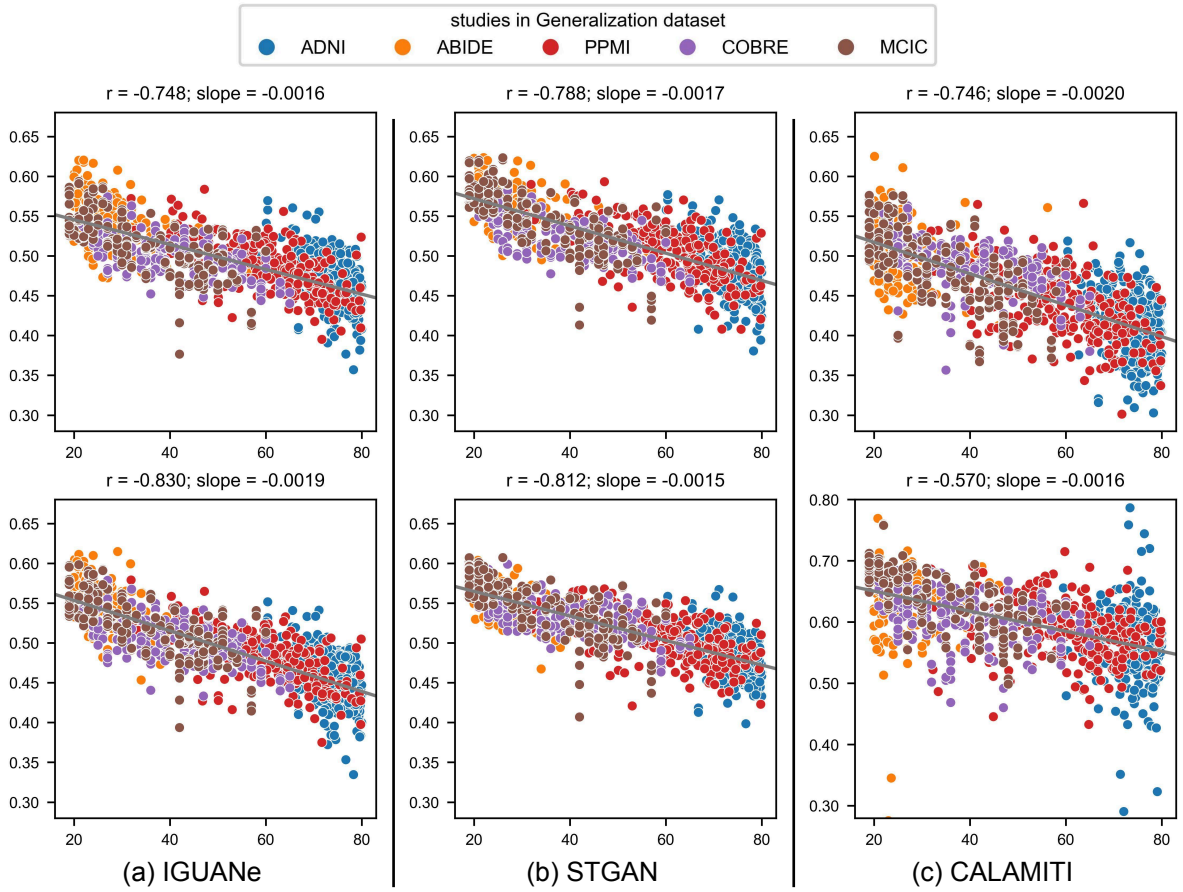
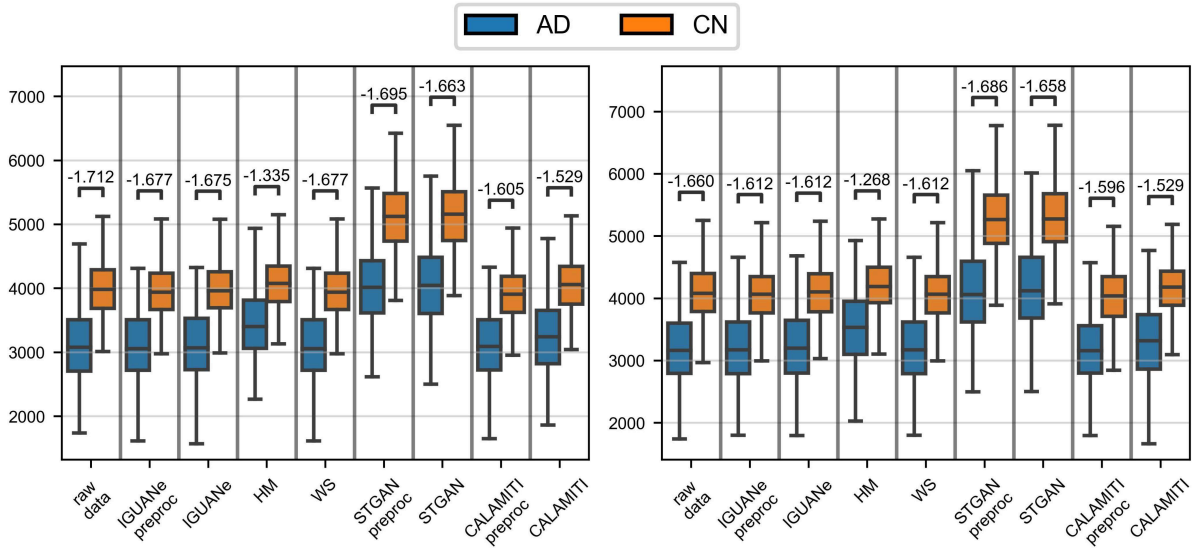


Figure 42 : Corrélation entre l'âge et les volumes de matière grise (MG) dans *Generalization dataset*. Les axes des X et des Y correspondent respectivement aux âges et aux volumes de MG (divisés par les volumes intracrâniens totaux), respectivement. Pour chaque modèle, les première et deuxième lignes montrent les données prétraitées et harmonisées, respectivement. La droite de régression linéaire est tracée sur chaque sous-figure.

5.3.4. Comparaison de volumes hippocampiques

Plusieurs observations peuvent être faites concernant les comparaisons entre les volumes hippocampiques des participants AD et ceux des CN (Figure 43). Tout d'abord, les mesures

étaient très élevées pour les images IRM prétraitées pour STGAN par rapport à toutes les autres mesures. Deuxièmement, IGUANe et WS ont mieux préservé la distinction entre les cas et les témoins que STGAN, CALAMITI et HM (en particulier ce dernier). Enfin, à l'exception de STGAN, qui a produit des résultats très différents, la taille de l'effet la plus importante a été obtenue sur les données brutes pour l'hippocampe droit et gauche.



(a) hippocampe gauche (b) hippocampe droit

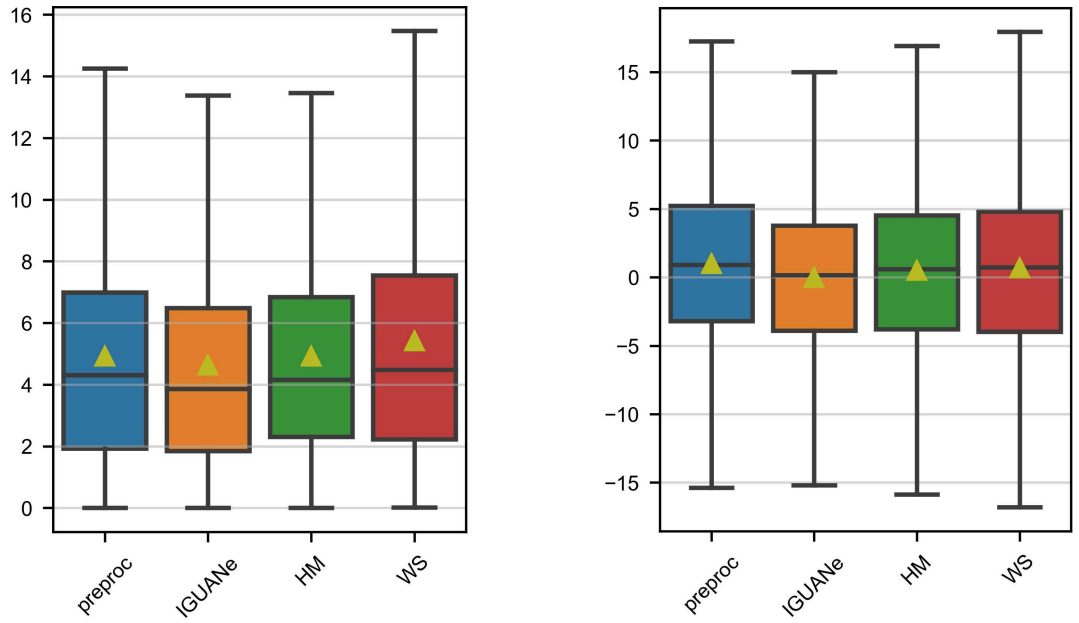
Figure 43 : Comparaisons des volumes hippocampiques estimés entre des participants sains (CN) et des participants avec la maladie d'Alzheimer (AD) dans *AD dataset*. Les volumes sont sur les axes des Y et sont exprimés en mm^3. Le terme *preproc* fait référence aux données obtenues après prétraitement pour l'approche d'harmonisation correspondante. Les d de Cohen comparant les groupes AD et CN sont indiqués au-dessus des diagrammes en boîte.

5.3.5. Prédiction d'âge

Dans la Figure 44a, on peut constater que la prédiction d'âge a été améliorée avec IGUANe (EAM de 4.92 à 4.63). En revanche, HM n'a pas vraiment modifié la précision (MAE = 4.92) et WS a entraîné une augmentation significative des erreurs (MAE = 5.42). On observe un léger motif de surestimation sur les données prétraitées et après HM et WS, tandis que les différences d'âge prédit étaient plus centrées autour de zéro après l'harmonisation IGUANe (Figure 44b).

5.3.6. Classification des participants sains et Alzheimer

Les résultats de la classification entre les participants CN et AD sont présentés dans le Tableau 6. Pour à la fois *AD_TEST* et *AD_GEN*, les meilleures performances ont été obtenues avec IGUANe. HM et WS ont réduit la précision dans *AD_TEST*, alors que dans *AD_GEN*, HM l'a légèrement augmentée et WS l'a maintenue inchangée.



(a) erreurs de prédiction (b) différences d'âge prédit

Figure 44 : Prédiction d'âge sur les images IRM de *Generalization dataset*. Le terme *preproc* fait référence aux données obtenues après prétraitement pour l'approche d'harmonisation correspondante. Les triangles indiquent les moyennes. Dans **(b)**, la différence d'âge prédit est calculée comme l'âge prédit moins l'âge réel.

Tableau 11 : Précisions dans la classification des participants sains et Alzheimer dans *AD dataset*.

Jeu de données / Méthode	preproc[1]	IGUANe	HM	WS
AD_TEST	0.843	**0.872**	0.809	0.834
AD_GEN	0.818	**0.840**	0.830	0.818

[1]*preproc* fait référence aux données après les prétraitements d'IGUANe.

5.4. Discussion

Dans cette étude, nous avons développé IGUANe, un modèle d'harmonisation inter-sites capable de traiter des images provenant de n'importe quel site. Après une phase d'entraînement avec un jeu de données comprenant 11 scanners pour un total de 4347 images cérébrales T1w, nous avons appliqué IGUANe à des données provenant d'autres études sans aucun fine-tuning. Nous avons mené différentes expériences pour évaluer la qualité des images générées et la préservation/renforcement des motifs biologiques. Des comparaisons avec deux techniques de normalisation d'intensité et deux méthodes de transfert de style démontrent la robustesse de la méthode proposée.

Comme StarGAN (Choi et al., 2018) et StarGAN v2 (Choi et al., 2020), IGUANe étend le modèle CycleGAN et dispose d'une procédure d'entraînement unifiée qui harmonise les images entre plusieurs domaines en parallèle. Cependant, contrairement à ces deux méthodes, l'apprentissage d'IGUANe repose sur des paires de domaines et permet l'utilisation de stratégies d'échantillonnage biaisé qui équilibrent les covariables biologiques,

évitant ainsi la suppression d'informations biologiques associées. Dans cette étude, nous avons mis en œuvre cette stratégie avec l'âge, ce qui nous a permis d'utiliser des ensembles d'entraînement avec des distributions d'âge très différentes (section 5.2.2.6). Le modèle de Gao et al. (2019), qui a en partie inspiré IGUANe, ne permet pas cela, car un seul discriminateur est utilisé pour le domaine de référence.

Les architectures de réseau dans IGUANe sont 3D et traitent des images de cerveau entières. À notre connaissance, à l'exception de celui que nous avons proposé dans la section 4, presque tous les modèles d'harmonisation non supervisée ont été conçus pour traiter des coupes ou des petits patchs de volume (section 2.5.3.4). Une autre originalité de la méthode est l'apprentissage résiduel. Pour l'harmonisation inter-sites, les modèles peuvent ainsi se concentrer sur l'apprentissage des modifications de contraste et non sur la reproduction de l'ensemble des informations anatomiques. Il convient également de noter que la définition d'une valeur neutre pour l'arrière-plan pendant l'entraînement améliore les images générées avec IGUANe ; Robinson et al. (2020) ont utilisé la même astuce dans leur modèle.

La visualisation des images IRM harmonisées (section 5.3.1) et les SSIMs obtenus avec les sujets voyageurs de SRPBS_TS (section 5.3.2) indiquent qu'IGUANe n'apporte pas de modifications substantielles aux images d'entrée. STGAN est moins conservateur et a réussi à augmenter légèrement les SSIMs dans SRPBS_TS. Cependant, les différences inter-sujets ont été beaucoup mieux préservées avec IGUANe (section 5.3.2), ce qui suggère que l'augmentation des SSIMs avec STGAN serait due à une sur-homogénéisation. Liu et al. (2023) ont montré une bonne préservation des variabilités inter-sujets avec STGAN par rapport à nous. La différence peut être due au fait que leur expérience était basée sur seulement 10 images IRM d'entrée, toutes acquises avec un scanner Siemens 3T de la même étude, qui était de plus incluse dans l'entraînement de STGAN. En accord avec des travaux antérieurs (sections 3.2.2.1 et 3.3), ces résultats illustrent donc les limites du SSIM en imagerie médicale.

En comparaison, l'analyse de la corrélation entre l'âge et le volume de MG a l'avantage d'évaluer les renforcements potentiels d'un motif spécifique de vieillissement cérébral. Une forte linéarité et une pente de régression élevée ont été obtenues avec IGUANe par rapport aux autres approches (section 5.3.3), ce qui suggère qu'IGUANe peut être utilisé comme une étape préliminaire dans les études basées sur des logiciels de segmentation automatique.

Nous avons insisté sur la préservation de motifs de vieillissement avec la prédiction d’âge. La prédiction d’âge a souvent été utilisée pour évaluer l'harmonisation en entraînant un modèle sur des images IRM d'un site et en l'appliquant sur des images IRM d'autres sites (Bashyam et al., 2022 ; Liu et al., 2023). Cependant, cette configuration n’est pas courante et les études récentes qui ont mis en place des modèles de prédiction d’âge disposaient de jeux d’entraînement multicentriques de grande taille (Bashyam et al., 2020 ; Cole et al., 2018 ; Gautherot et al., 2021), ce qui rend les modèles robustes aux effets de site. Ici, nous avons utilisé un tel jeu d’entraînement (section 5.2.4.4) et nous avons montré que, même dans cette configuration, les performances ont été améliorées avec l'harmonisation IGUANe, alors qu'elles sont restées similaires avec HM et diminuaient avec WS.

En plus de l'équilibrage des distributions d'âge, nous n'avons inclus que des participants sains dans le jeu d’entraînement afin d'éviter les effets confondants et la sur-correction. Cependant, les résultats montrent qu'IGUANe a non seulement maintenu les variabilités liées à l'âge dans les populations saines, mais aussi des motifs de la maladie d'Alzheimer. En effet, comme Liu et al. (2023) l'ont montré avec STGAN, IGUANe a été capable de

préserver les différences de volumes hippocampiques entre les participants CN et AD (section 5.3.4). Nous notons que les résultats de cette expérience suggèrent également que l'outil SynthSeg, qui a été conçu pour la segmentation de tout contraste (Billot et al., 2023), est mieux adapté pour traiter les images IRM avant le prétraitement et semble particulièrement biaisé avec les images IRM après le prétraitement STGAN. De plus, IGUANe a amélioré la classification CN/AD, que ce soit avec des données provenant de distributions similaires à l'ensemble d'entraînement du classifieur, ou avec des données acquises sur différentes machines de constructeurs inconnus (section 5.3.6). Par conséquent, ces résultats suggèrent qu'IGUANe se généralise bien non seulement aux images IRM de sites inconnus, mais aussi à des types de participants inconnus.

Nous avons comparé IGUANe à deux approches concurrentes récentes, STGAN et CALAMITI. D'autres méthodes récentes ont été proposées pour l'harmonisation des images IRM de sites non vus pendant l'entraînement.. Cependant, l'utilisation de ces modèles est rarement possible pour plusieurs raisons : complexité du code fourni et absence de modèle entraîné (Cackowski et al., 2023[12]), lien fourni sans code (Zuo et al., 2022[13]) ou aucun code fourni (Gao et al., 2019). Hu et al. (2023) ont également souligné ce problème pour les modèles d'harmonisation basés sur l'apprentissage profond. En outre, les mauvais résultats que nous avons obtenus avec CALAMITI (sections 5.3.3 et 5.3.4) pourraient également être liés à un problème de reproductibilité de la méthode, car il y a des ambiguïtés dans le prétraitement (section 5.2.3.2 et annexe 7.11). En effet, bien que Zuo et al. (2021b, 2021a) aient utilisé un fine-tuning pour adapter CALAMITI à de nouveaux sites, ils ont également montré des résultats corrects sans cela, ce qui n'est pas cohérent avec les nôtres.

Dans cette étude, nous nous sommes concentrés sur l'harmonisation d'images T1w, une séquence utilisée dans de nombreuses études de recherche et en pratique clinique. Le fait qu'IGUANe n'ait besoin que d'images T1w facilite son entraînement et son application. Il pourrait cependant être intéressant d'évaluer la capacité du modèle à harmoniser d'autres séquences. Des développements futurs pourraient également étudier les harmonisations parallèles de plusieurs séquences, ce qui serait possible avec IGUANe en ajoutant des canaux d'entrée/sortie dans les architectures de réseau. Dans cette voie, Dewey et al. (2019) ont constaté que la multimodalité a amélioré leur modèle d'harmonisation. Pour évaluer davantage la généralisation de notre approche, des expériences supplémentaires incluant d'autres pathologies pourraient également être menées. Par exemple, on peut se demander si l'harmonisation d'images IRM avec des lésions importantes nécessiterait un réentraînement avec des populations plus représentatives ou non afin d'éviter l'altération des caractéristiques des lésions.

5.5. Conclusion

Dans ce travail, nous introduisons IGUANe, une méthode générative non supervisée pour l'harmonisation inter-sites d'images IRM structurelles du cerveau. Nous fournissons un modèle entraîné avec un jeu d'entraînement multicentrique d'images T1w qui peut être facilement utilisé pour harmoniser des images de n'importe quel site. Le framework IGUANe comprend un entraînement antagoniste entre plusieurs modules d'apprentissage et a été conçu pour l'harmonisation sans sur-correction. Des expériences basées sur différentes cohortes de plusieurs études montrent que IGUANe améliore les motifs de vieillissement et

[12] lien fourni : https://github.com/nifm-gin/dl_generic accédé le 12/09/2023
[13] lien fourni : https://iacl.ece.jhu.edu/index.php?title=Resources accédé le 12/09/2023

les différences entre les participants sains et les participants atteints de la maladie d'Alzheimer. Elles montrent également la robustesse de l'approche par rapport à d'autres méthodes d'harmonisation. De futures études multicentriques pourraient utiliser IGUANe pour harmoniser leurs images, sans nécessiter de nouvelle phase d'entraînement.

6. Discussion générale

Ce travail de thèse a porté sur l'emploi de modèles génératifs non-supervisés pour l'harmonisation inter-sites d'images IRM structurelles du cerveau. Ce type de modèle n'est pas encore vraiment validé dans la littérature et de nombreux doutes persistent quant à leur pertinence en imagerie médicale. Nous avons donc d'abord présenté des revues bibliographiques ne se limitant pas à ses méthodes mais plus globalement à l'harmonisation rétrospective, en allant de prétraitements IRM classiques à des approches récentes d'apprentissage profond. Ces revues ont couvert à la fois la méthodologie pour l'harmonisation mais aussi la méthodologie pour son évaluation. Complétées par des expériences illustrant les potentielles limites d'approches couramment employées pour la validation de nouvelles méthodes, l'objectif était de proposer un point de vue documenté et critique sur les enjeux actuels dans ce domaine de recherche. Nous avons ensuite proposé deux modèles génératifs non-supervisés pour l'harmonisation inter-site d'images IRM structurelles, appliqués sur des images cérébrales T1w. Le premier fonctionne sur des paires de domaines et vise à exploiter des grands effectifs d'images IRM. Pour sa validation, l'accent a été mis sur la réduction de différences inter-domaines et la conservation d'informations IRM liées au vieillissement cérébral. Le deuxième modèle proposé est plus complexe méthodologiquement mais vise une plus grande souplesse pour l'application. En effet, la possibilité d'application à n'importe quelle image suite à l'entraînement est une amélioration majeure. L'évaluation de ce modèle s'est concentrée sur la conservation d'informations biologiques sur des images IRM acquises dans diverses études et avec des scanners très variés. Les résultats relatifs à ces deux modèles d'harmonisation mettent en avant leur efficacité mais surtout leur robustesse face à différents jeux de données et des applications subséquentes variables.

6.1. Les réseaux antagonistes génératifs pour l'harmonisation

Les modèles d'harmonisation proposés dans ce manuscrit (sections 4 et 5) reposent sur des GANs. Ce type de modèle a été beaucoup employé pour la synthèse d'images médicales ces dernières années (Kazeminia et al. 2020), notamment pour l'harmonisation (section 2.5.3). L'une des raisons du succès des GAN est qu'ils reposent sur un apprentissage non-supervisé avec très peu de contraintes.

Ces faibles contraintes peuvent toutefois être un inconvénient en imagerie médicale. Par exemple, Kazeminia et al. (2020) ont mis en avant le potentiel manque de contrôle sur l'algorithme qui est la cause d'un manque de fiabilité des images générées. Cette problématique est également liée à celle de l'interprétabilité qui fait généralement défaut avec l'apprentissage profond, et qui est également l'objet de nombreuses recherches en imagerie médicale (Salahuddin et al. 2022). Kazeminia et al. ont aussi évoqué d'autres problèmes courants dans l'utilisation des GAN comme l'instabilité de l'entraînement, un problème largement expérimenté avec les GAN (Creswell et al. 2018), et les évaluations souvent basées sur des métriques pixel-à-pixel requérant des vérités terrain. Singh et al. (2021) sont allés dans ce sens et selon eux, les GAN n'en sont qu'à leur début en imagerie médicale.

Le modèle CycleGAN ajoute de la contrainte à l'entraînement des GAN en les conditionnant avec une image d'entrée et en ajoutant une fonction de coût. Il a toutefois déjà été montré que CycleGAN peut modifier l'information pathologique en cas de déséquilibres

dans les jeux d'entraînement (Cohen et al. 2018), à tel point que certains chercheurs ne recommandent pas l'utilisation de ces méthodes pour l'harmonisation (Hu et al. 2023).

Malgré les limites potentielles mises en avant, de nombreuses études ont mis à profit ce type d'approche pour l'harmonisation (section 2.5.3.1). Ce fut également notre cas avec le 3D CycleGAN (section 4) et le générateur universel IGUANe (section 5). Afin de renforcer les contraintes sur l'entraînement de nos modèles et d'encourager la préservation d'informations anatomiques, nous avons opté pour plusieurs stratégies : renforcement des contraintes de consistance du cycle et d'identité, légère diminution du champ de réception des discriminateurs patchGAN, architectures U-net avec des connexions sautées, application de masques cérébraux pendant l'entraînement et apprentissage résiduel (uniquement pour IGUANe). L'attention apportée aux informations biologiques des données d'entraînement et la mise en place d'une stratégie d'échantillonnage prenant en compte d'éventuels déséquilibres a également été décisive pour éviter la sur-correction.

6.2. L'importance de l'évaluation

Afin de rendre compte au mieux de la fiabilité des modèles d'harmonisation, une attention particulière doit être portée sur leur évaluation. En effet, l'harmonisation étant devenue un domaine de recherche à part entière, les chercheurs doivent veiller à ne pas oublier les applications subséquentes. Un parallèle peut être fait avec la prédiction d'âge sur ce point. Nous avons ainsi choisi d'accorder une part presque aussi importante à la revue des méthodes d'évaluation (section 3.2) qu'à celle des modèles d'harmonisation (section 2).

Pour l'évaluation, une multitude de méthodes a été proposée (section 3.2). Nous avons vu que parmi elles, les métriques de similarité au niveau voxel sur des sujets voyageurs étaient les plus employées. Cependant, ces métriques ne sont pas forcément des bons indicateurs de la qualité de l'harmonisation (sections 3.2.2.1 et 3.3). Accorder moins d'importance à ce type de validation et privilégier davantage les applications subséquentes pourrait être une évolution positive de ce domaine de recherche. À ce titre, de nombreuses approches ont porté sur l'analyse de caractéristiques IRM, de motifs biologiques et sur des modèles prédictifs (section 3).

Parmi elles, celles qui rendent compte de l'évolution de motifs pathologiques dans les données IRM avec l'harmonisation sont particulièrement pertinentes car elles permettent d'entrevoir des applications concrètes, notamment dans des études multicentriques visant l'identification de biomarqueurs. Bien que de plus en plus de bases de données accessibles publiquement incluent des participants diagnostiqués pour certaines maladies (e.g. ADNI, PPMI, AIBL et OASIS), la mise en place d'évaluations quantifiant l'information pathologique et l'apport de l'harmonisation reste assez rare. L'un des principaux obstacles est les facteurs de confusion comme l'âge, le sexe et les différentes caractéristiques d'acquisition. Dans ce manuscrit, nous avons pu proposer des évaluations portant sur la maladie d'Alzheimer grâce aux bases de données ADNI et AIBL (sections 5.2.4.3 et 5.2.4.4). En effet, ces deux bases contiennent des images IRM de participants sains et de participants malades dans des tranches d'âge similaires, ce qui facilite l'isolation du motif pathologique.

6.3. La comparaison des méthodes

Étant donné le nombre important de méthodes d'harmonisation qui ont été proposées ces dernières années, il serait souhaitable de pouvoir les comparer afin de pouvoir déterminer

les plus efficaces, ou au moins les plus appropriées suivant l'application subséquente. Or, très peu d'études ont proposé ce genre de comparaison ou se limitent à des méthodes plus anciennes (Bashyam et al. 2022). Zuo et al. (2021b) ont comparé leur modèle d'harmonisation notamment à CycleGAN et à la méthode de Dewey et al. (2020). Peu de détails ont été donnés sur l'implémentation de CycleGAN utilisée par les auteurs et il est ainsi difficile d'en tirer des enseignements. La méthode de Dewey et al. est une méthode complexe d'apprentissage profond qui a été proposée par les mêmes auteurs mais pour lequel aucun code n'est accessible, rendant presque impossible la reproductibilité de l'expérience. La comparaison de Liu et al. (2023) est encore plus discutable puisqu'aucune information ni référence ne sont fournies concernant les méthodes comparées "cycleGAN" et "starGAN". Cackowski et al. (2023) ont également fait une comparaison d'harmonisation avec CycleGAN, sans fournir d'implémentation. Les auteurs ont de plus comparé de manière discutable leur modèle d'harmonisation à celui de Zuo et al. (2021a) (section 3.3.1). Nous pouvons aussi critiquer la comparaison que nous avons proposée entre notre CycleGAN 3D et un CycleGAN 2D (section 4.2.5). En effet, bien que nous ayons fourni plus de détails d'implémentation que dans les articles précédemment cités, d'autres choix auraient pu être faits pour la mise en œuvre du CycleGAN 2D.

Hu et al. (2023) ont insisté sur le problème de la comparabilité et de l'accès aux codes des méthodes d'harmonisation, en insistant sur les approches d'apprentissage profond. Selon eux, même quand un code est fourni, son utilisation est souvent difficile et pousse leurs utilisateurs à réimplémenter la méthode par eux-même. Hu et al. ont aussi mis en avant le paradoxe entre l'exclusion des détails d'implémentation dans les manuscrits publiés et l'importance que peuvent avoir ces détails sur les résultats. Pour pallier cette absence de reproductibilité, les auteurs encouragent notamment la réduction des dépendances d'implémentation, le suivi de pratiques d'ingénierie logicielle et la conteneurisation.

Dans notre étude présentant le modèle IGUANe (section 5), nous avons essayé de proposer une comparaison juste en reprenant les codes et les poids des modèles fournis par Liu et al. (2023) et Zuo et al. (2021b, 2021a). Le modèle IGUANe et les poids du générateur entraîné seront prochainement accessibles en ligne et les résultats présentés devraient être reproductibles.

6.4. Perspectives autour des travaux de thèse

Les deux modèles originaux présentés dans ce manuscrit reposent sur de la translation de domaines, avec des entraînements antagonistes incitant à l'homogénéisation des distributions entre les différents sites d'acquisition (ou scanners). Comme vu dans la section 2.5.3.3, beaucoup d'approches récentes ont été conçues à partir de transfert de style. Des futurs travaux pourraient essayer de déterminer le type d'approches le plus adapté suivant les données à disposition. On peut par exemple anticiper de meilleures performances avec notre modèle CycleGAN 3D (section 4) pour l'harmonisation de deux domaines pour lesquels de nombreuses images IRM sont à disposition. En revanche, l'harmonisation d'images réparties dans de nombreux petits domaines pourrait être plus efficace avec des approches de transfert de style déplaçant la focalisation sur le style des images IRM (Cackowski et al. 2023; Zuo et al. 2021b). De même, nous avons pu constater la robustesse du modèle IGUANe entraîné pour l'harmonisation d'images cérébrales T1w (section 4), mais la capacité de généralisation de cette approche complexe à d'autres types d'image sans bases de données de grande taille à disposition est une question ouverte.

Une autre perspective relative aux travaux de ce manuscrit concerne la méthode et l'architecture CycleGAN, qui repose notamment sur une contrainte de consistance du cycle au niveau voxel. Ce type de contraintes peut inciter les générateurs à préserver les structures mais peut être limité pour la synthèse des détails locaux dans l'image. Une pratique consiste alors à intégrer des fonctions de coût *perceptuelles* reposant sur les caractéristiques intermédiaires d'un discriminateur, censés extraire des informations de plus haut niveau dans l'image (Armanious et al. 2020). Une idée similaire est l'utilisation de modèles entraînés à des tâches auxiliaires (e.g. segmentation) pour l'extraction de caractéristiques pertinentes (Armanious et al. 2020; Yan et al. 2019). Il serait intéressant de mesurer l'apport de ces approches à nos modèles d'harmonisation, particulièrement pour IGUANe, qui a montré une robustesse à diverses applications mais aussi peut-être une sur-conservation des images d'entrée (e.g. section 5.3.1). Cela est notamment dû aux fortes régularisations de l'entraînement que nous avons mises en place pour éviter la sur-correction (section 5.2.2), et des fonctions de coûts portant sur des caractéristiques de plus haut niveau dans l'image pourraient permettre de relâcher ces contraintes.

6.5. Conclusion

Cette thèse porte sur l'harmonisation inter-sites d'images IRM structurelles du cerveau avec une focalisation particulière sur les modèles génératifs non-supervisés. Une revue approfondie de la littérature sur l'harmonisation rétrospective donne un large aperçu des méthodes utilisées pour harmoniser des données IRM et pour valider ces harmonisations. Les limites et les enjeux actuels du champ de recherche sont mis en avant. Deux modèles d'harmonisation d'images cérébrales T1w sont ensuite proposés. Le premier est conçu pour l'harmonisation de paires de domaines et les validations mises en place montrent sa capacité à réduire les différences inter-domaines tout en préservant des informations de vieillissement cérébral. Le deuxième a une procédure d'entraînement unifiant l'harmonisation d'un nombre arbitraire de domaines. Les validations proposées suggèrent qu'il peut harmoniser des images de sites non connus et renforcer des motifs liés à l'âge ou à la maladie d'Alzheimer. De futures études pourraient utiliser ces modèles sur d'autres cohortes, d'autres pathologies et/ou d'autres types d'image afin de rendre compte de leur capacité à généraliser à différentes applications.

7. Annexes

7.1. Sélection des images IRM dans les jeux de données indépendants de la section 4

7.1.1. IXI

Nous avons seulement utilisé les images *HH* et *Guys*. Les participants avec des données manquantes pour l'âge ou avec plusieurs informations contradictoires pour l'âge dans les métadonnées ont été exclus. Quatre participants de *HH* et quatre de *Guys* ont été exclus parce que le cerveau était trop grand le long de l'axe antéro-postérieur dans l'espace du MNI (> 192 voxels).

7.1.2. OASIS-3

Nous avons utilisé uniquement des images cérébrales T1w acquises avec un scanner *Biograph_mMR* ou *TrioTim* et pour lesquelles le CDR était 0. Les images avec des métadonnées manquantes pour l'âge ont été exclues. Ensuite, pour éviter des chevauchements, tous les participants présents dans les deux groupes ont été retirés du groupe *TrioTim*. Une image IRM a été exclue à cause d'une qualité pauvre (pas de cerveau sur l'image). Sept images IRM de *Biograph_mMR* et une de *TrioTim* ont été exclues parce que le cerveau était trop grand le long de l'axe antéro-postérieur dans l'espace du MNI (> 192 voxels).

7.1.3. NMorphCH

Toutes les images ont été incluses dans l'étude.

7.1.4. NKI-RS

Etant donné le besoin d'un jeu de données de participants jeunes (i.e. avec une distribution d'âge similaire à celle de NMorpCH), seules les images IRM de participants âgés entre 20 et 46 ans ont été incluses dans l'étude. Nous avons trouvé que les distributions d'intensité des images IRM de session "NFB3" étaient très différentes de celles des autres images et nous avons donc décidé de les retirer du jeu de données. Trois images IRM ont été exclues parce que le cerveau était trop grand le long de l'axe antéro-postérieur dans l'espace du MNI (> 192 voxels).

7.2. Détails de l'entraînement du modèle CycleGAN proposé

Nous avons adopté une fonction de coût des moindres carrés pour l'entraînement antagoniste (Mao et al. 2017) car nous l'avons trouvée plus adaptée que celle de Wasserstein (Gulrajani et al. 2017) aux calculs en précision mixte. Chaque entraînement consiste en 300 époques de 200 étapes. Un optimiseur Adam (Kingma et Ba 2017) avec un taux d'apprentissage à 0.0002 pour les 150 premières époques et réduit linéairement à 0 sur les 150 époques suivantes est utilisé. L'entraînement dure environ 20 heures sur GPU

NVIDIA Quadro RTX 6000 avec TensorFlow v2.9.1. Une data-augmentation consiste en une translation aléatoire (± 5 voxels) dans les trois plans orthogonaux.

7.3. Comparaison des distributions d'âge dans les expériences sites-appariés de la section 4

En utilisant un seuil de valeur p de 0.05, aucune des trois paires de jeux de données n'a de différence significative dans les distributions d'âge (Tableau S1).

Tableau S1 : Test U de Mann-Whitney entre les distributions d'âge dans les expériences sites-appariés de la section 4.

paire de jeux de données	Site1/Site2	Site3/Site4	Site5/Site6
valeur p [1]	0.0678	0.0678	0.2648

[1] Une correction de Benjamini-Hochberg est appliquée.

7.4. Distribution d'âge pour chaque jeu de données indépendant de la section 4

Les distributions sont illustrées dans la Figure S1.

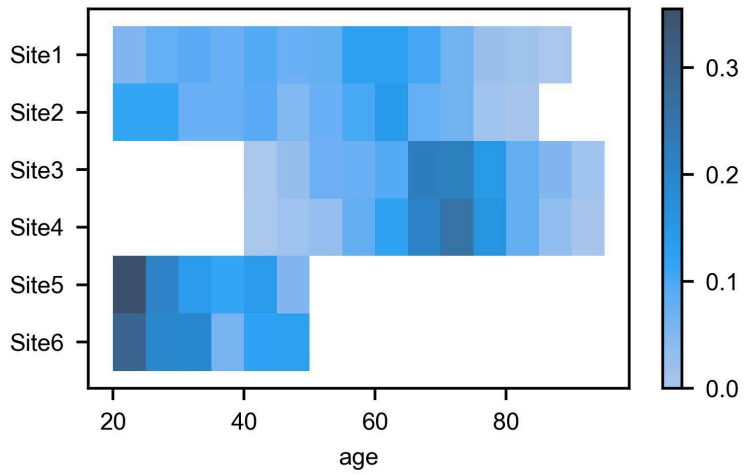


Figure S1 : Distribution d'âges pour chaque jeu de données indépendant de la section 4. Les proportions d'images IRM dans chaque tranche d'âge pour chaque jeu de données sont affichées.

7.5. Stratégie d'échantillonnage pour équilibrer les distributions d'âge

7.5.1. Procédure

Étant donné deux datasets A et B, nous avons défini une fonction qui prend la liste des âges dans A et dans B et un ensemble de tranches d'âge consécutives en entrée et qui retourne une probabilité d'échantillonnage pour chaque tranche.

Premièrement, le nombre d'images IRM dans chaque tranche d'âge et la distribution de probabilités correspondante sont calculés respectivement pour A et B. L'algorithme modifie ensuite les deux distributions itérativement et à chaque étape, les sous-étapes suivantes sont suivies :

1. Sélection de la tranche d'âge TA avec le plus petit nombre d'images (en considérant A et B séparément) parmi celles *non fixées*.
2. TA est marquée comme fixée.
3. La distribution de probabilités du jeu de données avec le plus d'images dans TA est mis à jour :
    a. La probabilité pour TA est mise à celle correspondante dans l'autre distribution.
    b. Les probabilités non fixées sont mises à jour proportionnellement en fonction de la modification de 3a, de telle sorte que la somme des probabilités est toujours égale à 1.

Après un nombre d'étapes égal au nombre de tranches d'âge moins 1, les deux distributions de probabilités sont égales et correspondent à la probabilité d'échantillonnage retournée par la fonction.

Le code de cette fonction est accessible dans un de nos répertoires en ligne (section 4.2.6).

7.5.2. Probabilités d'échantillonnage dans l'expérience multisite de la section 4

Les distributions de probabilités sont illustrées dans la Figure S2.

7.5.3. Probabilités d'échantillonnage pour l'entraînement d'IGUANe.

Les distributions de probabilités sont illustrées dans la Figure S3.

7.6. Détails d'entraînement de la prédiction d'âge dans la section 4

Tous les résultats dans ce manuscrit sont basés sur des modèles de prédiction d'âge entraînés sur 400 époques. Une data augmentation consiste en une translation aléatoire (± 5 voxels) dans les trois plans orthogonaux et une translation aléatoire (± 10 °) dans un plan sélectionné aléatoirement (rotation appliquée avec une probabilité 1/2). La rotation aléatoirement est appliquée une fois sur deux (fréquence aléatoire). Comme pour les procédures d'harmonisation, des calculs en précision mixte sont employés. La taille du batch

est fixée à 16. Nous avons utilisé un optimiseur Adam (Kingma et Ba 2017) avec un décroissement linéaire de 0.001 à 0.0001 pour le taux d'apprentissage.

Le code pour la prédiction d'âge est disponible en ligne (section 4.2.6).

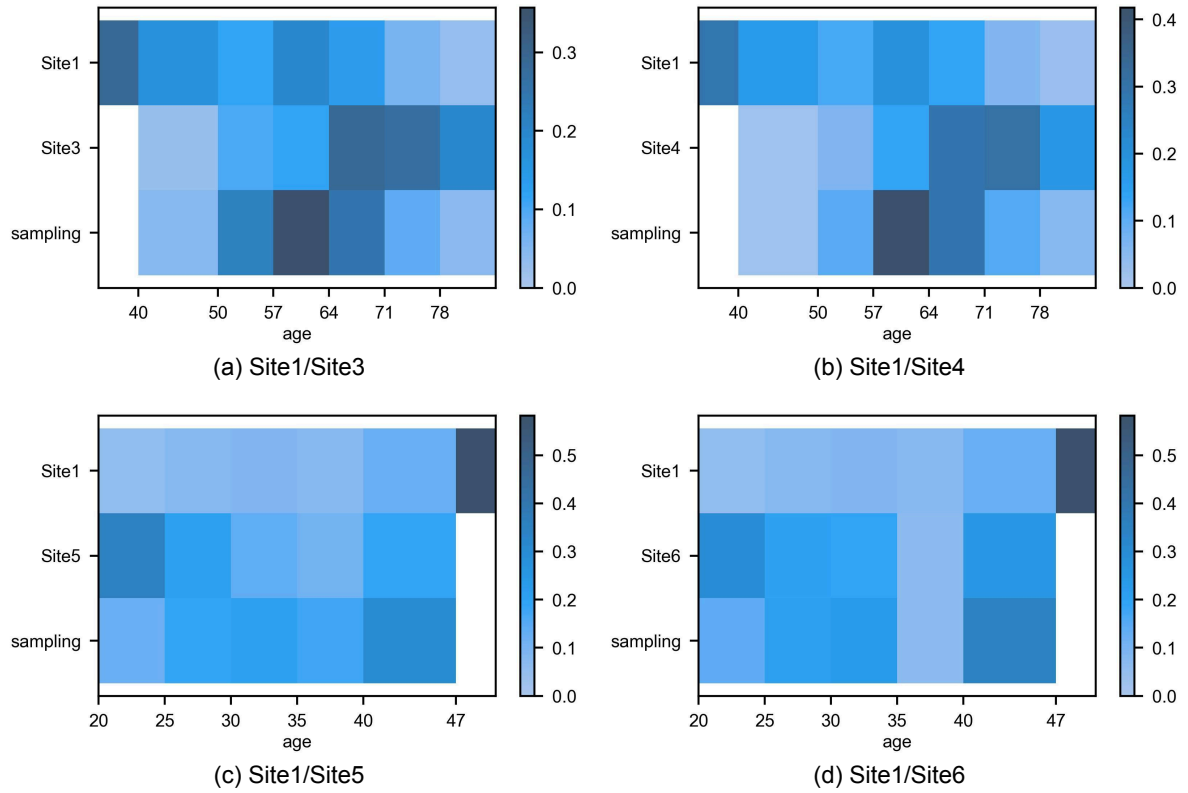


Figure S2 : Distributions de probabilités pour l'âge et pour l'échantillonnage pour chaque paire de jeu de données déséquilibrée en âge dans l'expérience multisite de la section 4. Les graduations sur l'axe des X indiquent les tranches d'âges utilisées pour l'échantillonnage. Les intensités de couleur indiquent les proportions pour chaque tranche d'âge.

7.7. CycleGAN 2D

Nous avons reproduit les architectures de réseau et la procédure d'entraînement décrits par Zhu et al. (2017) excepté pour le nombre de canaux d'entrée/sortie (1 au lieu de 3) et pour l'activation finale du générale que nous avons remplacé par celle que nous avons utilisé pour notre CycleGAN 3D (section 4.2.3.1). Pour chaque inférence, le masque original est appliqué sur le volume généré pour éviter les artefacts d'arrière-plan. Le nombre d'étapes d'entraînement par époque est 5000. Une data-augmentation consiste en une translation aléatoire (± 10 pixels) dans les deux plans orthogonaux.

7.8. Erreurs de prédiction d'âge par site avec le jeu d'entraînement multicentrique de la section 4

Les distributions d'erreurs sont illustrées dans la Figure S4.

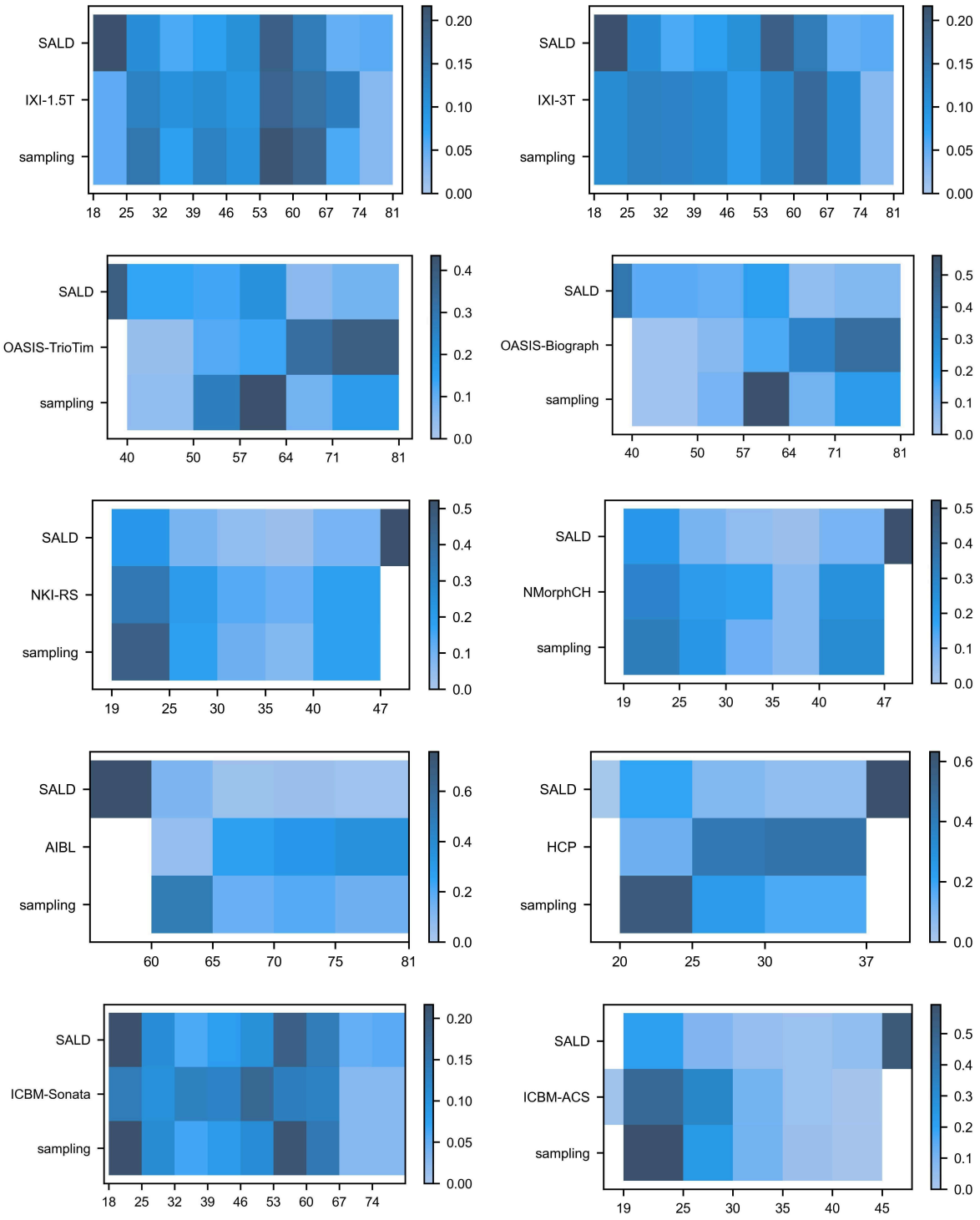


Figure S3 : Distributions de probabilités pour l'âge et pour l'échantillonnage pour l'entraînement d'IGUANe dans la section 5. Chaque sous-figure indique les distributions d'âge pour SALD et pour chaque site source ainsi que les probabilités d'échantillonnage utilisées pour l'entraînement. Les graduations sur l'axe des X indiquent les tranches d'âge utilisées pour l'échantillonnage. Les intensités de couleur indiquent les proportions pour chaque tranche d'âge.

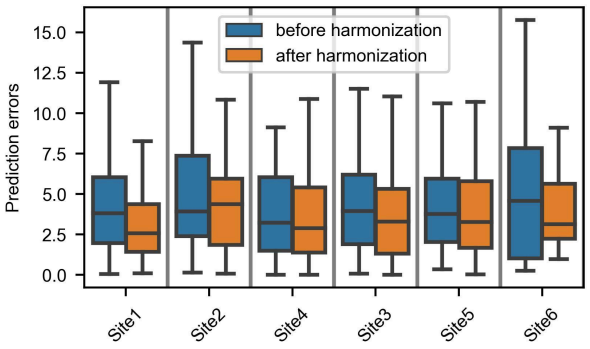


Figure S4 : Distributions des erreurs de prédiction d'âge par site avec jeu d'entraînement multicentrique dans l'expérience multisite de la section 4.

7.9. Résultats supplémentaires de prédiction d'âge sur des images de Site5 dans la section 4

Nous avons appliqué le modèle de prédiction d'âge entraîné avec les images de Site1 sur une version étendue du jeu de données de Site5 qui inclut des sujets plus vieux pas inclus dans l'étude principale. Nous avons harmonisé toutes ces images vers Site1 avec le modèle Site1/Site5 précédemment entraîné dans l'expérience multisite (section 4.2.4.1). L'EAM était de 6.54 années (DAPM : 2.07) sur les images originales et de 6.12 (DAPM : 1.02) sur les images harmonisées.

La Figure S5 montre les différences d'âge prédit en fonction des âges réels. En l'absence d'harmonisation, l'intersection entre la ligne de moyenne d'entraînement et la ligne de régression est clairement en dessous de 0 sur l'axe des Y, ce qui suggère que l'âge était sous-estimé. Ce n'est plus le cas après harmonisation où l'intersection est nettement plus proche de 0.

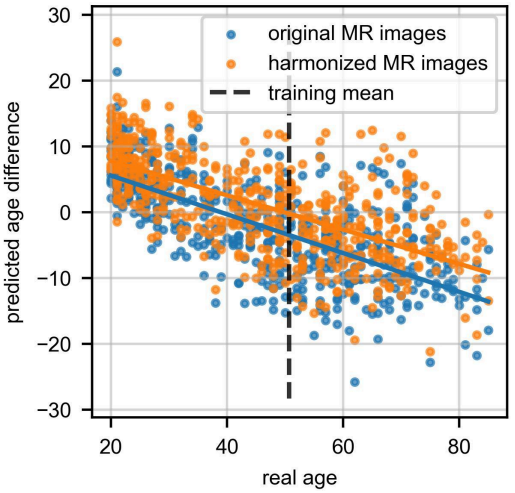


Figure S5 : Différence d'âge prédit en fonction de l'âge réel dans le jeu de données étendu de Site5 dans la section 4. La différence d'âge prédit est calculée comme l'âge prédit moins l'âge réel. Pour chaque ensemble d'images IRM, une ligne de régression des moindre carrés est tracée.

7.10. Détails supplémentaires de l'implémentation d'IGUANe

7.10.1. Fonctions de coût des générateurs

Une différence absolue par voxel moyenne est utilisée comme fonction de coût du cycle. Étant donné une image *x* et une image reconstruite après un cycle *x'*, la formule est la suivante : $L_{cyc} = |x-x'|_1 / N$ avec N le nombre de voxels.

La fonction de coût d'identité est calculée de la même manière. Étant donné une image *x* et l'image translatée vers le même domaine *x'*, la formule est : $L_{id} = |x-x'|_1 / N$.

La fonction de coût globale pour chaque générateur est ensuite calculée comme suit : $L = L_{adv} + \lambda * L_{cyc} + \lambda/2 * L_{id}$ avec L_{adv} la fonction de coût antagoniste. Nous avons fixé λ à 30.

7.10.2. Architecture des discriminateurs

Soient *x*, *y* et *z* des entiers. CxSyKz est un bloc Convolution-InstanceNormalization-LeakyReLU avec *x* filtres, des noyaux z^3 et des pas y^3. Le discriminateur est composé de 4 blocs consécutifs : C64S2K4-C128S2K4-C256S2K4-C512S1K3. Le premier bloc n'inclut pas de normalisation d'instance. Après le quatrième bloc, une convolution finale avec un pas de 1^3, génère le canal de sortie (activation linéaire).

7.10.3. Détails d'entraînement

Pour l'étude de la section 5, nous avons entraîné IGUANe sur 100 époques de 200 étapes (la procédure de chaque étape d'entraînement est donnée dans la section 5.2.2.5). Nous avons utilisé un optimiseur Adam (Kingma et Ba 2017) avec un décroissement linéaire de 0.002 à 0.0002 pour le taux d'apprentissage.

Nous avons implémenté une data-augmentation qui a consisté en une translation aléatoire (± 5 voxels) le long des trois axes orthogonaux et une rotation aléatoire (± 10 °) sur un plan orthogonal sélectionné aléatoirement (rotation appliqué avec une probabilité 1/2).

Pour la procédure de validation, nous avons sélectionné aléatoirement 44 participants pour chaque site source de la procédure de validation (44 correspond à l'ensemble avec le plus faible nombre de participants, i.e.NMorphCH). En utilisant la précision de prédiction du sexe et le coefficient de détermination (R^2) de la prédiction d'âge, nous avons défini la métrique suivante pour déterminer et sauvegarder le meilleur modèle : $0.75 * R^2 + 0.25 *$ précision. Des détails sur les modèles de prédiction utilisés dans la section 5 sont donnés dans l'annexe 7.12.

7.11. Normalisation d'intensité de CALAMITI pour la section 5

L'image T1w 3D fournie dans le répertoire en ligne de CALAMITI ne semblait pas avoir été normalisée avec WS de manière classique étant donné qu'elle ne contenait pas de valeurs négatives. Nous avons alors déterminé le masque de MB apparaissant normale du volume et avons calculé la moyenne et l'écart-type dans le masque (moyenne = 1016.9 ; écart-type = 7.8). Ensuite, afin d'implémenter les prétraitements CALAMITI, nous avons fait correspondre ces deux valeurs avec une mise à l'échelle et un décalage des intensités pour chaque image à normaliser (après les trois étapes de prétraitement expliquées en ligne).

7.12. Détails supplémentaires pour les modèles de prédiction de la section 5

Pour la régression d'âge et la classification CN/AD, nous avons reproduit la procédure de la section 4 (annexe 7.6). Le classifieur CN/AD a une activation finale sigmoid et a été entraîné avec une entropie croisée binaire au lieu d'une EAM.

Chaque image IRM normalisée avec HM a ses intensités mises à l'échelle et déplacées avec des constantes de telle sorte que les 1er et 99e percentiles des intensités cérébrales étaient -0.5 et 0.5, respectivement.

Chaque image normalisée avec WS a ses intensités mises à l'échelle et déplacées avec des constantes de telle sorte que la moyenne et l'écart-type de la MB apparaissant normale était 0.7 et 0.1, respectivement.

Références

Acquitter C., Piram L., Sabatini U., Gilhodes J., Moyal Cohen-Jonathan E., Ken S., et Lemasson B. 2022. « Radiomics-Based Detection of Radionecrosis Using Harmonized Multiparametric MRI ». Cancers 14(2):286. doi: 10.3390/cancers14020286.

Aine C. J., Bockholt H. J., Bustillo J. R., Cañive J. M., Caprihan A., Gasparovic C., Hanlon F. M., Houck J. M., Jung R. E., Lauriello J., Liu J., Mayer A. R., Perrone-Bizzozero N. I., Posse S., Stephen J. M., Turner J. A., Clark V. P., et Calhoun V. D. 2017. « Multimodal Neuroimaging in Schizophrenia: Description and Dissemination ». Neuroinformatics 15(4):343-64. doi: 10.1007/s12021-017-9338-9.

Alami Mejjati Y., Richardt C., Tompkin J., Cosker D., et Kim K. I. 2018. « Unsupervised Attention-guided Image-to-Image Translation ». in Advances in Neural Information Processing Systems. Vol. 31. Curran Associates, Inc.

Armanious K., Jiang C., Fischer M., Küstner T., Hepp T., Nikolaou K., Gatidis S., et Yang B. 2020. « MedGAN: Medical Image Translation Using GANs ». Computerized Medical Imaging and Graphics 79:101684. doi: 10.1016/j.compmedimag.2019.101684.

Basaia S., Agosta F., Wagner L., Canu E., Magnani G., Santangelo R., et Filippi M. 2019. « Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks ». NeuroImage: Clinical 21:101645. doi: 10.1016/j.nicl.2018.101645.

Bashyam V. M., Doshi J., Erus G., Srinivasan D., Abdulkadir A., Singh A., Habes M., Fan Y., Masters C. L., Maruff P., Zhuo C., Völzke H., Johnson S. C., Fripp J., Koutsouleris N., Satterthwaite T. D., Wolf D. H., Gur R. E., Gur R. C., Morris J. C., Albert M. S., Grabe H. J., Resnick S. M., Bryan N. R., Wittfeld K., Bülow R., Wolk D. A., Shou H., Nasrallah I. M., et Davatzikos C. 2022. « Deep Generative Medical Image Harmonization for Improving Cross-Site Generalization in Deep Learning Predictors ». Journal of Magnetic Resonance Imaging 55(3):908-16. doi: 10.1002/jmri.27908.

Bashyam V. M., Erus G., Doshi J., Habes M., Nasrallah I. M., Truelove-Hill M., Srinivasan D., Mamourian L., Pomponio R., Fan Y., Launer L. J., Masters C. L., Maruff P., Zhuo C., Völzke H., Johnson S. C., Fripp J., Koutsouleris N., Satterthwaite T. D., Wolf D., Gur R. E., Gur R. C., Morris J., Albert M. S., Grabe H. J., Resnick S., Bryan R. N., Wolk D. A., Shou H., et Davatzikos C. 2020. « MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide ». Brain 143(7):2312-24. doi: 10.1093/brain/awaa160.

Bayer J. M. M., Dinga R., Kia S. M., Kottaram A. R., Wolfers T., Lv J., Zalesky A., Schmaal L., et Marquand A. 2022. « Accommodating site variation in neuroimaging data using normative and hierarchical Bayesian models ». NeuroImage 264:119699. doi: 10.1016/j.neuroimage.2022.119699.

Beer J. C., Tustison N. J., Cook P. A., Davatzikos C., Sheline Y. I., Shinohara R. T., et Linn K. A. 2020. « Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data ». NeuroImage 220:117129. doi: 10.1016/j.neuroimage.2020.117129.

de Bel T., Bokhorst J.-M., van der Laak J., et Litjens G. 2021. « Residual cyclegan for robust domain transformation of histopathological tissue slides ». Medical Image Analysis 70:102004. doi: 10.1016/j.media.2021.102004.

Billot B., Greve D. N., Puonti O., Thielscher A., Van Leemput K., Fischl B., Dalca A. V., et Iglesias J. E. 2023. « SynthSeg: Segmentation of Brain MRI Scans of Any Contrast and Resolution without Retraining ». Medical Image Analysis 86:102789. doi: 10.1016/j.media.2023.102789.

Butler E. R., Chen A., Ramadan R., Le T. T., Ruparel K., Moore T. M., Satterthwaite T. D., Zhang F., Shou H., Gur R. C., Nichols T. E., et Shinohara R. T. 2021. « Pitfalls in brain age analyses ». Human Brain Mapping 42(13):4092-4101. doi: 10.1002/hbm.25533.

Cackowski S., Barbier E. L., Dojat M., et Christen T. 2023. « ImUnity: A Generalizable VAE-GAN Solution for Multicenter MR Image Harmonization ». Medical Image Analysis 102799. doi: 10.1016/j.media.2023.102799.
Cao S., Konz N., Duncan J., et Mazurowski M. A. 2023. « Deep Learning for Breast MRI Style Transfer with Limited Training Data ». Journal of Digital Imaging 36(2):666‑78. doi: 10.1007/s10278-022-00755-z.
Cha S.-H. 2008. « Taxonomy of Nominal Type Histogram Distance Measures ». MATH '.
Chang X., Cai X., Dan Y., Song Y., Lu Q., Yang G., et Nie S. 2022. « Self-Supervised Learning for Multi-Center Magnetic Resonance Imaging Harmonization without Traveling Phantoms ». Physics in Medicine & Biology 67(14):145004. doi: 10.1088/1361-6560/ac7b66.
Chauhan R., Ghanshala K. K., et Joshi R. C. 2018. « Convolutional Neural Network (CNN) for Image Detection and Recognition ». P. 278‑82 in 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC).
Chen A. A., Beer J. C., Tustison N. J., Cook P. A., Shinohara R. T., et Shou H. 2022. « Mitigating Site Effects in Covariance for Machine Learning in Neuroimaging Data ». Human Brain Mapping 43(4):1179‑95. doi: 10.1002/hbm.25688.
Chen J., Liu J., Calhoun V. D., Arias-Vasquez A., Zwiers M. P., Gupta C. N., Franke B., et Turner J. A. 2014. « Exploration of Scanning Effects in Multi-Site Structural MRI Studies ». Journal of Neuroscience Methods 230:37‑50. doi: 10.1016/j.jneumeth.2014.04.023.
Chen J., Sun Y., Fang Z., Lin W., Li G., et Wang L. 2021. « Harmonized Neonatal Brain MR Image Segmentation Model for Cross-Site Datasets ». Biomedical Signal Processing and Control 69:102810. doi: 10.1016/j.bspc.2021.102810.
Choi Y., Choi M., Kim M., Ha J.-W., Kim S., et Choo J. 2018. « StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation ». P. 8789‑97 in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.
Choi Y., Uh Y., Yoo J., et Ha J.-W. 2020. « StarGAN v2: Diverse Image Synthesis for Multiple Domains ». P. 8185‑94 in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
Cicchetti D. V. 1994. « Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology ». Psychological Assessment 6:284‑90. doi: 10.1037/1040-3590.6.4.284.
Clarke W. T., Mougin O., Driver I. D., Rua C., Morgan A. T., Asghar M., Clare S., Francis S., Wise R. G., Rodgers C. T., Carpenter A., Muir K., et Bowtell R. 2020. « Multi-site harmonization of 7 tesla MRI neuroimaging protocols ». NeuroImage 206:116335. doi: 10.1016/j.neuroimage.2019.116335.
Cohen J. P., Luck M., et Honari S. 2018. « Distribution Matching Losses Can Hallucinate Features in Medical Image Translation ». P. 529‑36 in Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Lecture Notes in Computer Science, édité par A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, et G. Fichtinger. Cham: Springer International Publishing.
Cole J. H., Poudel R. P. K., Tsagkrasoulis D., Caan M. W. A., Steves C., Spector T. D., et Montana G. 2017. « Predicting Brain Age with Deep Learning from Raw Imaging Data Results in a Reliable and Heritable Biomarker ». NeuroImage 163:115‑24. doi: 10.1016/j.neuroimage.2017.07.059.
Cole J. H., Ritchie S. J., Bastin M. E., Valdés Hernández M. C., Muñoz Maniega S., Royle N., Corley J., Pattie A., Harris S. E., Zhang Q., Wray N. R., Redmond P., Marioni R. E., Starr J. M., Cox S. R., Wardlaw J. M., Sharp D. J., et Deary I. J. 2018. « Brain Age Predicts Mortality ». Molecular Psychiatry 23(5):1385‑92. doi: 10.1038/mp.2017.62.
Creswell A., White T., Dumoulin V., Arulkumaran K., Sengupta B., et Bharath A. A. 2018. « Generative Adversarial Networks: An Overview ». IEEE Signal Processing Magazine 35(1):53‑65. doi: 10.1109/MSP.2017.2765202.
Dai X., Lei Y., Fu Y., Curran W. J., Liu T., Mao H., et Yang X. 2020. « Multimodal MRI

Synthesis Using Unified Generative Adversarial Networks ». Medical Physics 47(12):6343‑54. doi: 10.1002/mp.14539.
Dar S. U., Yurt M., Karacan L., Erdem A., Erdem E., et Çukur T. 2019. « Image Synthesis in Multi-Contrast MRI With Conditional Generative Adversarial Networks ». IEEE Transactions on Medical Imaging 38(10):2375‑88. doi: 10.1109/TMI.2019.2901750.
De Stefano N., Battaglini M., Pareto D., Cortese R., Zhang J., Oesingmann N., Prados F., Rocca M. A., Valsasina P., Vrenken H., Gandini Wheeler-Kingshott C. A. M., Filippi M., Barkhof F., et Rovira À. 2022. « MAGNIMS recommendations for harmonization of MRI data in MS multicenter studies ». NeuroImage: Clinical 34:102972. doi: 10.1016/j.nicl.2022.102972.
DeSilvio T., Moroianu S., Bhattacharya I., Seetharaman A., Sonn G., et Rusu M. 2021. « Intensity normalization of prostate MRIs using conditional generative adversarial networks for cancer detection ». P. 121‑26 in Medical Imaging 2021: Computer-Aided Diagnosis. Vol. 11597. SPIE.
Despotović I., Goossens B., et Philips W. 2015. « MRI Segmentation of the Human Brain: Challenges, Methods, and Applications ». Computational and Mathematical Methods in Medicine 2015:450341. doi: 10.1155/2015/450341.
Dewey B. E., Zhao C., Reinhold J. C., Carass A., Fitzgerald K. C., Sotirchos E. S., Saidha S., Oh J., Pham D. L., Calabresi P. A., van Zijl P. C. M., et Prince J. L. 2019. « DeepHarmony: A Deep Learning Approach to Contrast Harmonization across Scanner Changes ». Magnetic Resonance Imaging 64:160‑70. doi: 10.1016/j.mri.2019.05.041.
Dewey B. E., Zuo L., Carass A., He Y., Liu Y., Mowry E. M., Newsome S., Oh J., Calabresi P. A., et Prince J. L. 2020. « A Disentangled Latent Space for Cross-Site MRI Harmonization ». P. 720‑29 in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Lecture Notes in Computer Science, édité par A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, et L. Joskowicz. Cham: Springer International Publishing.
Dinsdale N. K., Jenkinson M., et Namburete A. I. L. 2021. « Deep Learning-Based Unlearning of Dataset Bias for MRI Harmonisation and Confound Removal ». NeuroImage 228:117689. doi: 10.1016/j.neuroimage.2020.117689.
Dosselmann R., et Yang X. D. 2011. « A Comprehensive Assessment of the Structural Similarity Index ». Signal, Image and Video Processing 5(1):81‑91. doi: 10.1007/s11760-009-0144-1.
Ellis K. A., Bush A. I., Darby D., De Fazio D., Foster J., Hudson P., Lautenschlager N. T., Lenzo N., Martins R. N., Maruff P., Masters C., Milner A., Pike K., Rowe C., Savage G., Szoeke C., Taddei K., Villemagne V., Woodward M., Ames D., et AIBL Research Group. 2009. « The Australian Imaging, Biomarkers and Lifestyle (AIBL) Study of Aging: Methodology and Baseline Characteristics of 1112 Individuals Recruited for a Longitudinal Study of Alzheimer's Disease ». International Psychogeriatrics 21(4):672‑87. doi: 10.1017/S1041610209009405.
Enriquez Calzada P. 2021. « Quantitative MR Inter-Scanner Harmonization Using Image Style Transfer ».
Erickson K. I., Leckie R. L., et Weinstein A. M. 2014. « Physical activity, fitness, and gray matter volume ». Neurobiology of aging 35 Suppl 2:S20‑28. doi: 10.1016/j.neurobiolaging.2014.03.034.
Esteban O., Birman D., Schaer M., Koyejo O. O., Poldrack R. A., et Gorgolewski K. J. 2017. « MRIQC: Advancing the Automatic Prediction of Image Quality in MRI from Unseen Sites ». PLOS ONE 12(9):e0184661. doi: 10.1371/journal.pone.0184661.
Fatania K., Clark A., Frood R., Scarsbrook A., Al-Qaisieh B., Currie S., et Nix M. 2022. « Harmonisation of scanner-dependent contrast variations in magnetic resonance imaging for radiation oncology, using style-blind auto-encoders ». Physics and Imaging in Radiation Oncology 22:115‑22. doi: 10.1016/j.phro.2022.05.005.
Fisher R. A. 1992. « Statistical Methods for Research Workers ». P. 66‑70 in Breakthroughs in Statistics: Methodology and Distribution, Springer Series in Statistics, édité par S.

Kotz et N. L. Johnson. New York, NY: Springer.
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., McInnis M., Phillips M. L., Trivedi M. H., Weissman M. M., et Shinohara R. T. 2018. « Harmonization of Cortical Thickness Measurements across Scanners and Sites ». NeuroImage 167:104‑20. doi: 10.1016/j.neuroimage.2017.11.024.
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., Schultz R. T., Verma R., et Shinohara R. T. 2017. « Harmonization of Multi-Site Diffusion Tensor Imaging Data ». NeuroImage 161:149‑70. doi: 10.1016/j.neuroimage.2017.08.047.
Fortin J.-P., Sweeney E. M., Muschelli J., Crainiceanu C. M., et Shinohara R. T. 2016. « Removing Inter-Subject Technical Variability in Magnetic Resonance Imaging Studies ». NeuroImage 132:198‑212. doi: 10.1016/j.neuroimage.2016.02.036.
Fuhai S., et Tang X. 2021. « Multi-modal MRI synthesization based on StarGAN ». P. 19‑22 in The Fourth International Symposium on Image Computing and Digital Medicine, ISICDM 2020. New York, NY, USA: Association for Computing Machinery.
Gao Y., Liu Y., Wang Y., Shi Z., et Yu J. 2019. « A Universal Intensity Standardization Method Based on a Many-to-One Weak-Paired Cycle Generative Adversarial Network for Magnetic Resonance Images ». IEEE Transactions on Medical Imaging 38(9):2059‑69. doi: 10.1109/TMI.2019.2894692.
Garcia-Dias R., Scarpazza C., Baecker L., Vieira S., Pinaya W. H. L., Corvin A., Redolfi A., Nelson B., Crespo-Facorro B., McDonald C., Tordesillas-Gutiérrez D., Cannon D., Mothersill D., Hernaus D., Morris D., Setien-Suero E., Donohoe G., Frisoni G., Tronchin G., Sato J., Marcelis M., Kempton M., van Haren N. E. M., Gruber O., McGorry P., Amminger P., McGuire P., Gong Q., Kahn R. S., Ayesa-Arriola R., van Amelsvoort T., Ortiz-García de la Foz V., Calhoun V., Cahn W., et Mechelli A. 2020. « Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners ». Neuroimage 220. doi: 10.1016/j.neuroimage.2020.117127.
Gautam P., Nuñez S. C., Narr K. L., Kan E. C., et Sowell E. R. 2014. « Effects of prenatal alcohol exposure on the development of white matter volume and change in executive function ». NeuroImage : Clinical 5:19‑27. doi: 10.1016/j.nicl.2014.05.010.
Gautherot M., Kuchcinski G., Bordier C., Sillaire A. R., Delbeuck X., Leroy M., Leclerc X., Pruvo J.-P., Pasquier F., et Lopes R. 2021. « Longitudinal Analysis of Brain-Predicted Age in Amnestic and Non-amnestic Sporadic Early-Onset Alzheimer's Disease ». Frontiers in Aging Neuroscience 13:729635. doi: 10.3389/fnagi.2021.729635.
Ge Y., Grossman R. I., Babb J. S., Rabin M. L., Mannon L. J., et Kolson D. L. 2002. « Age-Related Total Gray Matter and White Matter Changes in Normal Adult Brain. Part I: Volumetric MR Imaging Analysis ». American Journal of Neuroradiology 23(8):1327‑33.
Gebre R. K., Senjem M. L., Raghavan S., Schwarz C. G., Gunter J. L., Hofrenning E. I., Reid R. I., Kantarci K., Graff-Radford J., Knopman D. S., Petersen R. C., Jack C. R., et Vemuri P. 2023. « Cross–Scanner Harmonization Methods for Structural MRI May Need Further Work: A Comparison Study ». NeuroImage 269:119912. doi: 10.1016/j.neuroimage.2023.119912.
Gollub R. L., Shoemaker J. M., King M. D., White T., Ehrlich S., Sponheim S. R., Clark V. P., Turner J. A., Mueller B. A., Magnotta V., O'Leary D., Ho B. C., Brauns S., Manoach D. S., Seidman L., Bustillo J. R., Lauriello J., Bockholt J., Lim K. O., Rosen B. R., Schulz S. C., Calhoun V. D., et Andreasen N. C. 2013. « The MCIC collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia ». Neuroinformatics 11(3):367‑88. doi: 10.1007/s12021-013-9184-3.
Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., et Bengio Y. 2014. « Generative Adversarial Networks ».
Gourdeau D., Duchesne S., et Archambault L. 2022. « On the Proper Use of Structural Similarity for the Robust Evaluation of Medical Image Synthesis Models ». Medical

Physics 49(4):2462‑74. doi: 10.1002/mp.15514.
van Griethuysen J. J., Fedorov A., Parmar C., Hosny A., Aucoin N., Narayan V., Beets-Tan R. G., Fillion-Robin J.-C., Pieper S., et Aerts H. J. 2017. « Computational Radiomics System to Decode the Radiographic Phenotype ». Cancer research 77(21):e104‑7. doi: 10.1158/0008-5472.CAN-17-0339.
Grieve S. M., Korgaonkar M. S., Koslow S. H., Gordon E., et Williams L. M. 2013. « Widespread reductions in gray matter volume in depression ». NeuroImage : Clinical 3:332‑39. doi: 10.1016/j.nicl.2013.08.016.
Guan H., Liu Y., Yang E., Yap P.-T., Shen D., et Liu M. 2021. « Multi-Site MRI Harmonization via Attention-Guided Deep Domain Adaptation for Brain Disorder Identification ». Medical Image Analysis 71:102076. doi: 10.1016/j.media.2021.102076.
Gulrajani I., Ahmed F., Arjovsky M., Dumoulin V., et Courville A. C. 2017. « Improved Training of Wasserstein GANs ». in Advances in Neural Information Processing Systems. Vol. 30. Curran Associates, Inc.
Hardaha S., Edla D. R., et Parne S. R. 2023. « A Survey on Convolutional Neural Networks for MRI Analysis ». Wireless Personal Communications 128(2):1065‑85. doi: 10.1007/s11277-022-09989-0.
Hawco C., Viviano J. D., Chavez S., Dickie E. W., Calarco N., Kochunov P., Argyelan M., Turner J. A., Malhotra A. K., Buchanan R. W., et Voineskos A. N. 2018. « A longitudinal human phantom reliability study of multi-center T1-weighted, DTI, and resting state fMRI data ». Psychiatry Research: Neuroimaging 282:134‑42. doi: 10.1016/j.pscychresns.2018.06.004.
He K., Zhang X., Ren S., et Sun J. 2016. « Deep Residual Learning for Image Recognition ». P. 770‑78 in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Hedman A. M., van Haren N. E. M., Schnack H. G., Kahn R. S., et Hulshoff Pol H. E. 2012. « Human Brain Changes across the Life Span: A Review of 56 Longitudinal Magnetic Resonance Imaging Studies ». Human Brain Mapping 33(8):1987‑2002. doi: 10.1002/hbm.21334.
Hognon C., Tixier F., Gallinato O., Colin T., Visvikis D., et Jaouen V. 2019. « Standardization of Multicentric Image Datasets with Generative Adversarial Networks ». in IEEE Nuclear Science Symposium and Medical Imaging Conference 2019. Manchester, United Kingdom.
Hu F., Chen A. A., Horng H., Bashyam V., Davatzikos C., Alexander-Bloch A., Li M., Shou H., Satterthwaite T. D., Yu M., et Shinohara R. T. 2023. « Image Harmonization: A Review of Statistical and Deep Learning Methods for Removing Batch Effects and Evaluation Metrics for Effective Harmonization ». NeuroImage 274:120125. doi: 10.1016/j.neuroimage.2023.120125.
Huang X., et Belongie S. 2017. « Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization ». P. 1510‑19 in 2017 IEEE International Conference on Computer Vision (ICCV).
Huo Y., Xu Z., Xiong Y., Aboud K., Parvathaneni P., Bao S., Bermudez C., Resnick S. M., Cutting L. E., et Landman B. A. 2019. « 3D whole brain segmentation using spatially localized atlas network tiles ». NeuroImage 194:105‑19. doi: 10.1016/j.neuroimage.2019.03.041.
Iqbal S., Ghani Khan M. U., Saba T., Mehmood Z., Javaid N., Rehman A., et Abbasi R. 2019. « Deep Learning Model Integrating Features and Novel Classifiers Fusion for Brain Tumor Segmentation ». Microscopy Research and Technique 82(8):1302‑15. doi: 10.1002/jemt.23281.
Isensee F., Schell M., Pflueger I., Brugnara G., Bonekamp D., Neuberger U., Wick A., Schlemmer H.-P., Heiland S., Wick W., Bendszus M., Maier-Hein K. H., et Kickingereder P. 2019. « Automated Brain Extraction of Multisequence MRI Using Artificial Neural Networks ». Human Brain Mapping 40(17):4952‑64. doi: 10.1002/hbm.24750.
Išgum I., Benders M. J. N. L., Avants B., Cardoso M. J., Counsell S. J., Gomez E. F., Gui L.,

Hűppi P. S., Kersbergen K. J., Makropoulos A., Melbourne A., Moeskops P., Mol C. P., Kuklisova-Murgasova M., Rueckert D., Schnabel J. A., Srhoj-Egekher V., Wu J., Wang S., de Vries L. S., et Viergever M. A. 2015. « Evaluation of automatic neonatal brain segmentation algorithms: The NeoBrainS12 challenge ». Medical Image Analysis 20(1):135‑51. doi: 10.1016/j.media.2014.11.001.
Isola P., Zhu J.-Y., Zhou T., et Efros A. A. 2017. « Image-to-Image Translation with Conditional Adversarial Networks ». P. 5967‑76 in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Jenkinson M., Bannister P., Brady M., et Smith S. 2002. « Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images ». NeuroImage 17(2):825‑41. doi: 10.1006/nimg.2002.1132.
Jonsson B. A., Bjornsdottir G., Thorgeirsson T. E., Ellingsen L. M., Walters G. B., Gudbjartsson D. F., Stefansson H., Stefansson K., et Ulfarsson M. O. 2019. « Brain Age Prediction Using Deep Learning Uncovers Associated Sequence Variants ». Nature Communications 10(1):5409. doi: 10.1038/s41467-019-13163-9.
Jovicich J., Czanner S., Greve D., Haley E., van der Kouwe A., Gollub R., Kennedy D., Schmitt F., Brown G., MacFall J., Fischl B., et Dale A. 2006. « Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data ». NeuroImage 30(2):436‑43. doi: 10.1016/j.neuroimage.2005.09.046.
Kazeminia S., Baur C., Kuijper A., van Ginneken B., Navab N., Albarqouni S., et Mukhopadhyay A. 2020. « GANs for medical image analysis ». Artificial Intelligence in Medicine 109:101938. doi: 10.1016/j.artmed.2020.101938.
Kieselmann J. P., Fuller C. D., Gurney-Champion O. J., et Oelfke U. 2021. « Cross-Modality Deep Learning: Contouring of MRI Data from Annotated CT Data Only ». Medical Physics 48(4):1673‑84. doi: 10.1002/mp.14619.
Kingma D. P., et Ba J. 2017. « Adam: A Method for Stochastic Optimization ».
Kruggel F., Turner J., et Muftuler L. T. 2010. « Impact of Scanner Hardware and Imaging Protocol on Image Quality and Compartment Volume Precision in the ADNI Cohort ». NeuroImage 49(3):2123‑33. doi: 10.1016/j.neuroimage.2009.11.006.
LaMontagne P. J., Benzinger T. L., Morris J. C., Keefe S., Hornbeck R., Xiong C., Grant E., Hassenstab J., Moulder K., Vlassenko A. G., Raichle M. E., Cruchaga C., et Marcus D. 2019. « OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease ». 2019.12.13.19014902.
Li W., Li Y., Qin W., Liang X., Xu J., Xiong J., et Xie Y. 2020. « Magnetic resonance image (MRI) synthesis from brain computed tomography (CT) images based on deep learning methods for magnetic resonance (MR)-guided radiotherapy ». Quantitative Imaging in Medicine and Surgery 10(6):1223‑36. doi: 10.21037/qims-19-885.
Li Y., Ammari S., Balleyguier C., Lassau N., et Chouzenoux E. 2021. « Impact of Preprocessing and Harmonization Methods on the Removal of Scanner Effects in Brain MRI Radiomic Features ». Cancers 13(12):3000. doi: 10.3390/cancers13123000.
Liu M., Maiti P., Thomopoulos S., Zhu A., Chai Y., Kim H., et Jahanshad N. 2021. « Style Transfer Using Generative Adversarial Networks for Multi-Site MRI Harmonization ». P. 313‑22 in Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Lecture Notes in Computer Science, édité par M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, et C. Essert. Cham: Springer International Publishing.
Liu M., Zhu A. H., Maiti P., Thomopoulos S. I., Gadewar S., Chai Y., Kim H., et Jahanshad N. 2023. « Style Transfer Generative Adversarial Networks to Harmonize Multisite MRI to a Single Reference Image to Avoid Overcorrection ». Human Brain Mapping 44(14):4875‑92. doi: 10.1002/hbm.26422.
Liu S., et Yap P.-T. 2021. « Learning Multi-Site Harmonization of Magnetic Resonance Images Without Traveling Human Phantoms ». arXiv:2110.00041 [cs, eess].
Lundervold A. S., et Lundervold A. 2019. « An Overview of Deep Learning in Medical Imaging Focusing on MRI ». Zeitschrift Für Medizinische Physik 29(2):102‑27. doi:

10.1016/j.zemedi.2018.11.002.
Maaten L. van der, et Hinton G. 2008. « Visualizing Data using t-SNE ». Journal of Machine Learning Research 9(86):2579‑2605.
MacLullich A. M. J., Wardlaw J. M., Ferguson K. J., Starr J. M., Seckl J. R., et Deary I. J. 2004. « Enlarged Perivascular Spaces Are Associated with Cognitive Function in Healthy Elderly Men ». Journal of Neurology, Neurosurgery & Psychiatry 75(11):1519‑23. doi: 10.1136/jnnp.2003.030858.
Maikusa N., Zhu Y., Uematsu A., Yamashita A., Saotome K., Okada N., Kasai K., Okanoya K., Yamashita O., Tanaka S. C., et Koike S. 2021. « Comparison of traveling‐subject and ComBat harmonization methods for assessing structural brain characteristics ». Human Brain Mapping 42(16):5278‑87. doi: 10.1002/hbm.25615.
Mallya Y., J V., S V. M., Venugopal V. K., et Mahajan V. 2019. « Automatic delineation of anterior and posterior cruciate ligaments by combining deep learning and deformable atlas based segmentation ». P. 471‑77 in Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging. Vol. 10953. SPIE.
Manjón J. V., et Coupé P. 2016. « volBrain: An Online MRI Brain Volumetry System ». Frontiers in Neuroinformatics 10. doi: 10.3389/fninf.2016.00030.
Mao X., Li Q., Xie H., Lau R. Y. K., Wang Z., et Smolley S. P. 2017. « Least Squares Generative Adversarial Networks ». P. 2813‑21 in 2017 IEEE International Conference on Computer Vision (ICCV).
Marek K., Chowdhury S., Siderowf A., Lasch S., Coffey C. S., Caspell-Garcia C., Simuni T., Jennings D., Tanner C. M., Trojanowski J. Q., Shaw L. M., Seibyl J., Schuff N., Singleton A., Kieburtz K., Toga A. W., Mollenhauer B., Galasko D., Chahine L. M., Weintraub D., Foroud T., Tosun-Turgut D., Poston K., Arnedo V., Frasier M., Sherer T., et Initiative the P. P. M. 2018. « The Parkinson's Progression Markers Initiative (PPMI) – Establishing a PD Biomarker Cohort ». Annals of Clinical and Translational Neurology 5(12):1460‑77. doi: 10.1002/acn3.644.
Micikevicius P., Narang S., Alben J., Diamos G., Elsen E., Garcia D., Ginsburg B., Houston M., Kuchaiev O., Venkatesh G., et Wu H. 2018. « Mixed Precision Training ».
Miskin N., Patel H., Franceschi A. M., Ades-Aron B., Le A., Damadian B. E., Stanton C., Serulle Y., Golomb J., Gonen O., Rusinek H., et George A. E. 2017. « Diagnosis of Normal-Pressure Hydrocephalus: Use of Traditional Measures in the Era of Volumetric MR Imaging ». Radiology 285(1):197‑205. doi: 10.1148/radiol.2017161216.
Modanwal G., Vellal A., et Mazurowski M. A. 2021. « Normalization of Breast MRIs Using Cycle-Consistent Generative Adversarial Networks ». Computer Methods and Programs in Biomedicine 208:106225. doi: 10.1016/j.cmpb.2021.106225.
Naser M. A., et Deen M. J. 2020. « Brain Tumor Segmentation and Grading of Lower-Grade Glioma Using Deep Learning in MRI Images ». Computers in Biology and Medicine 121:103758. doi: 10.1016/j.compbiomed.2020.103758.
Nguyen H., Morris R. W., Harris A. W., Korgoankar M. S., et Ramos F. 2018. « Correcting differences in multi-site neuroimaging data using Generative Adversarial Networks ». arXiv:1803.09375 [cs].
Nie D., Trullo R., Lian J., Wang L., Petitjean C., Ruan S., Wang Q., et Shen D. 2018. « Medical Image Synthesis with Deep Convolutional Adversarial Networks ». IEEE Transactions on Biomedical Engineering 65(12):2720‑30. doi: 10.1109/TBME.2018.2814538.
Nooner K., Colcombe S., Tobe R., Mennes M., Benedict M., Moreno A., Panek L., Brown S., Zavitz S., Li Q., Sikka S., Gutman D., Bangaru S., Schlachter R. T., Kamiel S., Anwar A., Hinz C., Kaplan M., Rachlin A., Adelsberg S., Cheung B., Khanuja R., Yan C., Craddock C., Calhoun V., Courtney W., King M., Wood D., Cox C., Kelly C., DiMartino A., Petkova E., Reiss P., Duan N., Thompsen D., Biswal B., Coffey B., Hoptman M., Javitt D., Pomara N., Sidtis J., Koplewicz H., Castellanos F., Leventhal B., et Milham M. 2012. « The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry ». Frontiers in Neuroscience 6. doi:

10.3389/fnins.2012.00152.
Nyul L. G., Udupa J. K., et Zhang X. 2000. « New variants of a method of MRI scale standardization ». IEEE Transactions on Medical Imaging 19(2):143‑50. doi: 10.1109/42.836373.
Odena A., Dumoulin V., et Olah C. 2016. « Deconvolution and Checkerboard Artifacts ». Distill 1(10):e3. doi: 10.23915/distill.00003.
Palladino J. A., Fernandez Slezak D., et Ferrante E. 2020. « Unsupervised domain adaptation via CycleGAN for white matter hyperintensity segmentation in multicenter MR images ». P. 14 in 16th International Symposium on Medical Information Processing and Analysis, édité par J. Brieva, N. Lepore, E. Romero Castro, et M. G. Linguraru. Lima, Peru: SPIE.
Palubinskas G. 2014. « Mystery behind similarity measures mse and SSIM ». P. 575‑79 in 2014 IEEE International Conference on Image Processing (ICIP).
Pambrun J.-F., et Noumeir R. 2015. « Limitations of the SSIM quality metric in the context of diagnostic imaging ». P. 2960‑63 in 2015 IEEE International Conference on Image Processing (ICIP).
Pasquier F., Leys D., Weerts J. G. E., Mounier-Vehier F., Barkhof F., et Scheltens P. 1996. « Inter-and Intraobserver Reproducibility of Cerebral Atrophy Assessment on MRI Scans with Hemispheric Infarcts ». European Neurology 36(5):268‑72. doi: 10.1159/000117270.
Pomponio R., Erus G., Habes M., Doshi J., Srinivasan D., Mamourian E., Bashyam V., Nasrallah I. M., Satterthwaite T. D., Fan Y., Launer L. J., Masters C. L., Maruff P., Zhuo C., Völzke H., Johnson S. C., Fripp J., Koutsouleris N., Wolf D. H., Gur R., Gur R., Morris J., Albert M. S., Grabe H. J., Resnick S. M., Bryan R. N., Wolk D. A., Shinohara R. T., Shou H., et Davatzikos C. 2020. « Harmonization of Large MRI Datasets for the Analysis of Brain Imaging Patterns throughout the Lifespan ». NeuroImage 208:116450. doi: 10.1016/j.neuroimage.2019.116450.
Qin Z., Liu Z., Zhu P., et Ling W. 2022. « Style transfer in conditional GANs for cross-modality synthesis of brain magnetic resonance images ». Computers in Biology and Medicine 148:105928. doi: 10.1016/j.compbiomed.2022.105928.
Qu L., Zhang Y., Wang S., Yap P.-T., et Shen D. 2020. « Synthesized 7T MRI from 3T MRI via Deep Learning in Spatial and Wavelet Domains ». Medical Image Analysis 62:101663. doi: 10.1016/j.media.2020.101663.
Radua J., Vieta E., Shinohara R., Kochunov P., Quidé Y., Green M. J., Weickert C. S., Weickert T., Bruggemann J., Kircher T., Nenadić I., Cairns M. J., Seal M., Schall U., Henskens F., Fullerton J. M., Mowry B., Pantelis C., Lenroot R., Cropley V., Loughland C., Scott R., Wolf D., Satterthwaite T. D., Tan Y., Sim K., Piras F., Spalletta G., Banaj N., Pomarol-Clotet E., Solanes A., Albajes-Eizagirre A., Canales-Rodríguez E. J., Sarro S., Di Giorgio A., Bertolino A., Stäblein M., Oertel V., Knöchel C., Borgwardt S., du Plessis S., Yun J.-Y., Kwon J. S., Dannlowski U., Hahn T., Grotegerd D., Alloza C., Arango C., Janssen J., Díaz-Caneja C., Jiang W., Calhoun V., Ehrlich S., Yang K., Cascella N. G., Takayanagi Y., Sawa A., Tomyshev A., Lebedeva I., Kaleda V., Kirschner M., Hoschl C., Tomecek D., Skoch A., van Amelsvoort T., Bakker G., James A., Preda A., Weideman A., Stein D. J., Howells F., Uhlmann A., Temmingh H., López-Jaramillo C., Díaz-Zuluaga A., Fortea L., Martinez-Heras E., Solana E., Llufriu S., Jahanshad N., Thompson P., Turner J., van Erp T., Glahn D., Pearlson G., Hong E., Krug A., Carr V., Tooney P., Cooper G., Rasser P., Michie P., Catts S., Gur R., Gur R., Yang F., Fan F., Chen J., Guo H., Tan S., Wang Z., Xiang H., Piras F., Assogna F., Salvador R., McKenna P., Bonvino A., King M., Kaiser S., Nguyen D., et Pineda-Zapata J. 2020. « Increased Power by Harmonizing Structural MRI Site Differences with the ComBat Batch Adjustment Method in ENIGMA ». NeuroImage 218:116956. doi: 10.1016/j.neuroimage.2020.116956.
Rathore S., Habes M., Iftikhar M. A., Shacklett A., et Davatzikos C. 2017. « A Review on Neuroimaging-Based Classification Studies and Associated Feature Extraction

Methods for Alzheimer's Disease and Its Prodromal Stages ». NeuroImage 155:530‑48. doi: 10.1016/j.neuroimage.2017.03.057.
Ravano V., Démonet J.-F., Damian D., Meuli R., Piredda G. F., Huelnhagen T., Maréchal B., Thiran J.-P., Kober T., et Richiardi J. 2022. « Neuroimaging Harmonization Using cGANs: Image Similarity Metrics Poorly Predict Cross-Protocol Volumetric Consistency ». P. 83‑92 in Machine Learning in Clinical Neuroimaging, Lecture Notes in Computer Science, édité par A. Abdulkadir, D. R. Bathula, N. C. Dvornek, M. Habes, S. M. Kia, V. Kumar, et T. Wolfers. Cham: Springer Nature Switzerland.
Reig S., Sánchez-González J., Arango C., Castro J., González-Pinto A., Ortuño F., Crespo-Facorro B., Bargalló N., et Desco M. 2009. « Assessment of the Increase in Variability When Combining Volumetric Data from Different Scanners ». Human Brain Mapping 30(2):355‑68. doi: 10.1002/hbm.20511.
Richter S., Winzeck S., Correia M. M., Kornaropoulos E. N., Manktelow A., Outtrim J., Chatfield D., Posti J. P., Tenovuo O., Williams G. B., Menon D. K., et Newcombe V. F. J. 2022. « Validation of Cross-Sectional and Longitudinal ComBat Harmonization Methods for Magnetic Resonance Imaging Data on a Travelling Subject Cohort ». Neuroimage: Reports 2(4):100136. doi: 10.1016/j.ynirp.2022.100136.
Robinson R., Dou Q., Coelho de Castro D., Kamnitsas K., de Groot M., Summers R. M., Rueckert D., et Glocker B. 2020. « Image-Level Harmonization of Multi-Site Data Using Image-and-Spatial Transformer Networks ». P. 710‑19 in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Lecture Notes in Computer Science, édité par A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, et L. Joskowicz. Cham: Springer International Publishing.
Robitaille N., Mouiha A., Crépeault B., Valdivia F., et Duchesne S. 2012. « Tissue-Based MRI Intensity Standardization: Application to Multicentric Datasets ». International Journal of Biomedical Imaging 2012:e347120. doi: 10.1155/2012/347120.
Sabuncu M. R., et Konukoglu E. 2015. « Clinical prediction from structural brain MRI scans: A large-scale empirical study ». Neuroinformatics 13(1):31‑46. doi: 10.1007/s12021-014-9238-1.
Sajedi H., et Pardakhti N. 2019. « Age Prediction Based on Brain MRI Image: A Survey ». Journal of Medical Systems 43(8):279. doi: 10.1007/s10916-019-1401-7.
Salahuddin Z., Woodruff H. C., Chatterjee A., et Lambin P. 2022. « Transparency of deep neural networks for medical image analysis: A review of interpretability methods ». Computers in Biology and Medicine 140:105111. doi: 10.1016/j.compbiomed.2021.105111.
Scheltens P., Launer L. J., Barkhof F., Weinstein H. C., et van Gool W. A. 1995. « Visual Assessment of Medial Temporal Lobe Atrophy on Magnetic Resonance Imaging: Interobserver Reliability ». Journal of Neurology 242(9):557‑60. doi: 10.1007/BF00868807.
Schmitter D., Roche A., Maréchal B., Ribes D., Abdulkadir A., Bach-Cuadra M., Daducci A., Granziera C., Klöppel S., Maeder P., Meuli R., et Krueger G. 2015. « An evaluation of volume-based morphometry for prediction of mild cognitive impairment and Alzheimer's disease ». NeuroImage: Clinical 7:7‑17. doi: 10.1016/j.nicl.2014.11.001.
Schuff N., Woerner N., Boreta L., Kornfield T., Shaw L. M., Trojanowski J. Q., Thompson P. M., Jack C. R. Jr, Weiner M. W., et the Alzheimer's; Disease Neuroimaging Initiative. 2009. « MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers ». Brain 132(4):1067‑77. doi: 10.1093/brain/awp007.
Shah M., Xiao Y., Subbanna N., Francis S., Arnold D. L., Collins D. L., et Arbel T. 2011. « Evaluating Intensity Normalization on MRIs of Human Brain with Multiple Sclerosis ». Medical Image Analysis 15(2):267‑82. doi: 10.1016/j.media.2010.12.003.
Shinohara R. T., Oh J., Nair G., Calabresi P. A., Davatzikos C., Doshi J., Henry R. G., Kim G., Linn K. A., Papinutto N., Pelletier D., Pham D. L., Reich D. S., Rooney W., Roy S., Stern W., Tummala S., Yousuf F., Zhu A., Sicotte N. L., et Bakshi R. 2017. « Volumetric Analysis from a Harmonized Multisite Brain MRI Study of a Single Subject

with Multiple Sclerosis ». American Journal of Neuroradiology 38(8):1501‑9. doi: 10.3174/ajnr.A5254.
Shinohara R. T., Sweeney E. M., Goldsmith J., Shiee N., Mateen F. J., Calabresi P. A., Jarso S., Pham D. L., Reich D. S., et Crainiceanu C. M. 2014. « Statistical Normalization Techniques for Magnetic Resonance Imaging ». NeuroImage: Clinical 6:9‑19. doi: 10.1016/j.nicl.2014.08.008.
Shrivastava A., Pfister T., Tuzel O., Susskind J., Wang W., et Webb R. 2017. « Learning from Simulated and Unsupervised Images through Adversarial Training ». P. 2242‑51 in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Sinha S., Thomopoulos S. I., Lam P., Muir A., et Thompson P. M. 2021. « Alzheimer's disease classification accuracy is Improved by MRI harmonization based on attention-guided generative adversarial networks ». P. 180‑89 in 17th International Symposium on Medical Information Processing and Analysis. Vol. 12088. SPIE.
Sled J. G., Zijdenbos A. P., et Evans A. C. 1998. « A nonparametric method for automatic correction of intensity nonuniformity in MRI data ». IEEE Transactions on Medical Imaging 17(1):87‑97. doi: 10.1109/42.668698.
Smith S. M. 2002. « Fast Robust Automated Brain Extraction ». Human Brain Mapping 17(3):143‑55. doi: 10.1002/hbm.10062.
Srivastava N., Hinton G., Krizhevsky A., Sutskever I., et Salakhutdinov R. 2014. « Dropout: A Simple Way to Prevent Neural Networks from Overfitting ». Journal of Machine Learning Research 15(56):1929‑58.
Steiger J. H. 1980. « Tests for comparing elements of a correlation matrix ». Psychological Bulletin 87(2):245‑51. doi: 10.1037/0033-2909.87.2.245.
Takao H., Hayashi N., et Ohtomo K. 2011. « Effect of Scanner in Longitudinal Studies of Brain Volume Changes ». Journal of Magnetic Resonance Imaging 34(2):438‑44. doi: 10.1002/jmri.22636.
Tan H. H. G., Westeneng H.-J., Nitert A. D., van Veenhuijzen K., Meier J. M., van der Burgh H. K., van Zandvoort M. J. E., van Es M. A., Veldink J. H., et van den Berg L. H. 2022. « MRI Clustering Reveals Three ALS Subtypes With Unique Neurodegeneration Patterns ». Annals of Neurology 92(6):1030‑45. doi: 10.1002/ana.26488.
Tanaka S. C., Yamashita A., Yahata N., Itahashi T., Lisi G., Yamada T., Ichikawa N., Takamura M., Yoshihara Y., Kunimatsu A., Okada N., Hashimoto R., Okada G., Sakai Y., Morimoto J., Narumoto J., Shimada Y., Mano H., Yoshida W., Seymour B., Shimizu T., Hosomi K., Saitoh Y., Kasai K., Kato N., Takahashi H., Okamoto Y., Yamashita O., Kawato M., et Imamizu H. 2021. « A Multi-Site, Multi-Disorder Resting-State Magnetic Resonance Image Database ». Scientific Data 8(1):227. doi: 10.1038/s41597-021-01004-8.
Tang H., Xu D., Sebe N., et Yan Y. 2019. « Attention-Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation ». P. 1‑8 in 2019 International Joint Conference on Neural Networks (IJCNN).
Tang Y., Chen D., et Li X. 2021. « Dimensionality Reduction Methods for Brain Imaging Data Analysis ». ACM Computing Surveys 54(4):87:1-87:36. doi: 10.1145/3448302.
Tian D., Zeng Z., Sun X., Tong Q., Li H., He H., Gao J.-H., He Y., et Xia M. 2022. « A Deep Learning-Based Multisite Neuroimage Harmonization Framework Established with a Traveling-Subject Dataset ». NeuroImage 257:119297. doi: 10.1016/j.neuroimage.2022.119297.
Tong Q., He H., Gong T., Li C., Liang P., Qian T., Sun Y., Ding Q., Li K., et Zhong J. 2020. « Multicenter Dataset of Multi-Shell Diffusion MRI in Healthy Traveling Adults with Identical Settings ». Scientific Data 7(1):157. doi: 10.1038/s41597-020-0493-8.
Torbati M. E., Tudorascu D. L., Minhas D. S., Maillard P., DeCarli C. S., et Jae Hwang S. 2021. « Multi-scanner Harmonization of Paired Neuroimaging Data via Structure Preserving Embedding Learning ». P. 3277‑86 in 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).
Tustison N. J., Avants B. B., Cook P. A., Zheng Y., Egan A., Yushkevich P. A., et Gee J. C.

2010. « N4ITK: Improved N3 Bias Correction ». IEEE Transactions on Medical Imaging 29(6):1310-20. doi: 10.1109/TMI.2010.2046908.
Ulyanov D., Vedaldi A., et Lempitsky V. 2017. « Instance Normalization: The Missing Ingredient for Fast Stylization ».
Vieira S., Pinaya W. H. L., et Mechelli A. 2017. « Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications ». Neuroscience & Biobehavioral Reviews 74:58-75. doi: 10.1016/j.neubiorev.2017.01.002.
Wachinger C., Rieckmann A., et Pölsterl S. 2021. « Detect and Correct Bias in Multi-Site Neuroimaging Datasets ». Medical Image Analysis 67:101879. doi: 10.1016/j.media.2020.101879.
Wang R., Chaudhari P., et Davatzikos C. 2022. « Embracing the Disharmony in Medical Imaging: A Simple and Effective Framework for Domain Adaptation ». Medical Image Analysis 76:102309. doi: 10.1016/j.media.2021.102309.
Wang T., et Lin Y. 2018. « CycleGAN with Better Cycles ».
Wang Z., Bovik A. C., Sheikh H. R., et Simoncelli E. P. 2004. « Image quality assessment: from error visibility to structural similarity ». IEEE Transactions on Image Processing 13(4):600-612. doi: 10.1109/TIP.2003.819861.
Wardlaw J. M., Smith E. E., Biessels G. J., Cordonnier C., Fazekas F., Frayne R., Lindley R. I., O'Brien J. T., Barkhof F., Benavente O. R., Black S. E., Brayne C., Breteler M., Chabriat H., DeCarli C., de Leeuw F.-E., Doubal F., Duering M., Fox N. C., Greenberg S., Hachinski V., Kilimann I., Mok V., Oostenbrugge R. van, Pantoni L., Speck O., Stephan B. C. M., Teipel S., Viswanathan A., Werring D., Chen C., Smith C., van Buchem M., Norrving B., Gorelick P. B., et Dichgans M. 2013. « Neuroimaging Standards for Research into Small Vessel Disease and Its Contribution to Ageing and Neurodegeneration ». The Lancet Neurology 12(8):822-38. doi: 10.1016/S1474-4422(13)70124-8.
Watanabe M., Liao J., Jara H., et Sakai O. 2013. « Multispectral Quantitative MR Imaging of the Human Brain: Lifetime Age-related Effects ». Radiographics : a review publication of the Radiological Society of North America, Inc 33:1305-19. doi: 10.1148/rg.335125212.
Wei D., Zhuang K., Ai L., Chen Q., Yang W., Liu W., Wang K., Sun J., et Qiu J. 2018. « Structural and Functional Brain Scans from the Cross-Sectional Southwest University Adult Lifespan Dataset ». Scientific Data 5(1):180134. doi: 10.1038/sdata.2018.134.
Welander P., Karlsson S., et Eklund A. 2018. « Generative Adversarial Networks for Image-to-Image Translation on Multi-Contrast MR Images - A Comparison of CycleGAN and UNIT ». arXiv:1806.07777 [cs].
Wolterink J. M., Dinkla A. M., Savenije M. H. F., Seevinck P. R., Berg C. A. T. van den, et Isgum I. 2017. « Deep MR to CT Synthesis using Unpaired Data ». arXiv:1708.01155 [cs].
Wrobel J., Martin M. L., Bakshi R., Calabresi P. A., Elliot M., Roalf D., Gur R. C., Gur R. E., Henry R. G., Nair G., Oh J., Papinutto N., Pelletier D., Reich D. S., Rooney W. D., Satterthwaite T. D., Stern W., Prabhakaran K., Sicotte N. L., Shinohara R. T., et Goldsmith J. 2020. « Intensity warping for multisite MRI harmonization ». NeuroImage 223:117242. doi: 10.1016/j.neuroimage.2020.117242.
Wu S., Li G., Deng L., Liu L., Wu D., Xie Y., et Shi L. 2019. « L1 -Norm Batch Normalization for Efficient Training of Deep Neural Networks ». IEEE Transactions on Neural Networks and Learning Systems 30(7):2043-51. doi: 10.1109/TNNLS.2018.2876179.
Xiang L., Li Y., Lin W., Wang Q., et Shen D. 2018. « Unpaired Deep Cross-Modality Synthesis with Fast Training ». P. 155-64 in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Lecture Notes in Computer Science, édité par D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. S. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, et A. Madabhushi. Cham: Springer International Publishing.

Yan W., Wang Y., Gu S., Huang L., Yan F., Xia L., et Tao Q. 2019. « The Domain Shift Problem of Medical Image Segmentation and Vendor-Adaptation by Unet-GAN ». P. 623‑31 in Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Lecture Notes in Computer Science, édité par D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, et A. Khan. Cham: Springer International Publishing.

Yi Z., Zhang H., Tan P., et Gong M. 2017. « DualGAN: Unsupervised Dual Learning for Image-to-Image Translation ». P. 2868‑76 in 2017 IEEE International Conference on Computer Vision (ICCV).

Yu M., Linn K. A., Cook P. A., Phillips M. L., McInnis M., Fava M., Trivedi M. H., Weissman M. M., Shinohara R. T., et Sheline Y. I. 2018. « Statistical Harmonization Corrects Site Effects in Functional Connectivity Measurements from Multi-Site fMRI Data ». Human Brain Mapping 39(11):4213‑27. doi: 10.1002/hbm.24241.

Zech J. R., Badgeley M. A., Liu M., Costa A. B., Titano J. J., et Oermann E. K. 2018. « Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study ». PLOS Medicine 15(11):e1002683. doi: 10.1371/journal.pmed.1002683.

Zhang Y., Brady M., et Smith S. 2001. « Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm ». IEEE Transactions on Medical Imaging 20(1):45‑57. doi: 10.1109/42.906424.

Zhao M., Wang L., Chen J., Nie D., Cong Y., Ahmad S., Ho A., Yuan P., Fung S. H., Deng H. H., Xia J., et Shen D. 2018. « Craniomaxillofacial Bony Structures Segmentation from MRI with Deep-Supervision Adversarial Learning ». P. 720‑27 in Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Lecture Notes in Computer Science, édité par A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, et G. Fichtinger. Cham: Springer International Publishing.

Zhong J., Wang Y., Li J., Xue X., Liu S., Wang M., Gao X., Wang Q., Yang J., et Li X. 2020. « Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: application to neonatal white matter development ». BioMedical Engineering OnLine 19(1):4. doi: 10.1186/s12938-020-0748-9.

Zhu J.-Y., Park T., Isola P., et Efros A. A. 2017. « Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks ». P. 2242‑51 in 2017 IEEE International Conference on Computer Vision (ICCV).

Zuo L., Dewey B. E., Carass A., Liu Y., He Y., Calabresi P. A., et Prince J. L. 2021a. « Information-Based Disentangled Representation Learning for Unsupervised MR Harmonization ». P. 346‑59 in Information Processing in Medical Imaging, Lecture Notes in Computer Science, édité par A. Feragen, S. Sommer, J. Schnabel, et M. Nielsen. Cham: Springer International Publishing.

Zuo L., Dewey B. E., Liu Y., He Y., Newsome S. D., Mowry E. M., Resnick S. M., Prince J. L., et Carass A. 2021b. « Unsupervised MR Harmonization by Learning Disentangled Representations Using Information Bottleneck Theory ». NeuroImage 243:118569. doi: 10.1016/j.neuroimage.2021.118569.

Zuo L., Liu Y., Xue Y., Han S., Bilgel M., Resnick S. M., Prince J. L., et Carass A. 2022. « Disentangling a Single MR Modality ». P. 54‑63 in Data Augmentation, Labelling, and Imperfections, Lecture Notes in Computer Science, édité par H. V. Nguyen, S. X. Huang, et Y. Xue. Cham: Springer Nature Switzerland.