Transfer learning from synthetic to routine clinical data for motion artefact detection in brain T1-weighted MRI

Sophie Loizillon, Simona Bottani, Aurélien Maire, Sebastian Ströer, Didier Dormont, Olivier Colliot, Ninon Burgos

Transfer learning from synthetic to routine clinical data for motion artefact detection in brain T1-weighted MRI

Sophie Loizillon[a], Simona Bottani[a], Aurélien Maire[b], Sebastian Ströer[c], Didier Dormont[a,c], Olivier Colliot[a], Ninon Burgos[a], and the APPRIMAGE Study Group[1]

[a]Sorbonne Université, Institut du Cerveau – Paris Brain Institute, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, Paris, France
[b]AP-HP, WIND department, Paris, France
[c]AP-HP, Hôpital Pitié-Salpêtrière, DMU DIAMENT, Dep. of Neuroradiology, Paris, France

ABSTRACT

Clinical data warehouses (CDWs) contain the medical data of millions of patients and represent a great opportunity to develop computational tools. MRIs are particularly sensitive to patient movements during image acquisition, which will result in artefacts (blurring, ghosting and ringing) in the reconstructed image. As a result, a significant number of MRIs in CDWs are unusable because corrupted by these artefacts. Since their manual detection is impossible due to the number of scans, it is necessary to develop a tool to automatically exclude images with motion in order to fully exploit CDWs. In this paper, we propose a CNN for the automatic detection of motion in 3D T1-weighted brain MRI. Our transfer learning approach, based on synthetic motion generation, consists of two steps: a pre-training on research data using synthetic motion, followed by a fine-tuning step to generalise our pre-trained model to clinical data, relying on the manual labelling of 5500 images. The objectives were both (1) to be able to exclude images with severe motion, (2) to detect mild motion artefacts. Our approach achieved excellent accuracy for the first objective with a balanced accuracy nearly similar to that of the annotators (balanced accuracy>80%). However, for the second objective, the performance was weaker and substantially lower than that of human raters. Overall, our framework will be useful to take advantage of CDWs in medical imaging and to highlight the importance of a clinical validation of models trained on research data.

Keywords: Quality Control, Clinical Data Warehouse, Deep Learning, Transfer Learning, Motion, MRI

1. INTRODUCTION

Recently, hospitals have created clinical data warehouses (CDWs) which gather medical images from thousands to millions of patients.[1] These resources represent an exceptional opportunity to develop computational tools.[2] In contrast to research datasets where acquisition protocols are well standardised, the quality of CDW images is highly heterogeneous. Images come from different hospitals over several decades and diverse machines were used with no homogenisation on the acquisition parameters.[3] Therefore, quality control (QC) is a fundamental first step before developing any machine learning project on a CDW. Lately, we have developed a framework for the automatic QC of T1-weighted (T1w) brain MRIs for a CDW based on noise, motion and contrast labels. While we obtained good results for the classification of images which are not proper 3D T1w brain MRI and for recognising low quality images, we encountered difficulties in detecting motion in the images.[4]

MRIs are sensitive to motion induced by patient movement during the acquisition process. As they require a long acquisition time, subjects are more likely to move during the examination, which causes artefacts (blurring, ringing, ghosting or signal loss) in the reconstructed image.[5] Previously, we found that 25% of MRIs in the CDW were totally unusable, and almost a third had a very low quality especially due to motion.[4] A study conducted in a single hospital showed that the prevalence of repeating an MRI examination due to the presence of motion was up to 20% of all the acquisitions.[6] Beyond the cost that this represents for hospitals, these studies are

Further author information: (Send correspondence to Sophie Loizillon)
A.A.A.: E-mail: sophie.loizillon@gmail.com

highlighting the fact that many images present in the PACS are corrupted by motion. Therefore, it is important to be able to automatically exclude such images before conducting any study on a CDW.

As motion quantification is a complex problem, particularly due to its sensitivity to many cofactors such as contrast, we suffer from a lack of dataset with reliable quantitative ground truths. Thus, some studies were based on synthetic motion to detect motion artefacts in a controlled way.[7–10] Despite the excellent results claimed in the literature, only few papers have attempted to validate their performance on data with real motion. Even when they did, their test sets were extremely limited and only composed of research data.[8,10] It is thus unclear how they would perform on clinical routine data.

In this paper, we propose a transfer learning framework for the automatic detection of motion artefacts in 3D T1w brain MRI using a CDW. We generated synthetic motion in MR images of research databases to train a CNN classifier which was validated on synthetic and real motion artefacts. Our model was then generalised to clinical data with an effective transfer learning technique using 5000 manually labelled MRIs from a CDW.

2. MATERIALS AND METHODS

2.1 Dataset description

Three publicly available research datasets were used to train a CNN with synthetic motion artefacts and one routine clinical dataset was exploited for transfer learning and validation.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) study is a multisite study of elderly individuals with normal cognition, mild cognitive impairment, or Alzheimer's disease.[11] The ADNI-1 phase included T1w MRIs acquired on 1.5 T scanners from different manufacturers (GE, Siemens, and Philips). Part of the metadata, the IPMOTION score indicates the absence of motion (0), or the presence of mild (1), moderate (2) or severe (3) motion artefacts in MRIs. This score was used to select motion free T1w MRI. We also created a small test set with MRI corrupted by severe, moderate or mild motion based on the IPMOTION score and the comments section. None of the ADNI images were acquired with gadolinium injection. We also used data from the MSSEG MICCAI challenge, which aim is to segment multiple sclerosis lesions and includes 53 patients across 4 different sites,[12] and from the Montreal Neurological Institute's Brain Images of Tumors (MNI BITE) database, which includes pre and postoperative MR images acquired from 13 brain tumour patients with different subtypes of gliomas.[13] We only considered the 3D T1w with gadolinium injection from these two datasets.

The clinical routine data comes from a large CDW containing all the T1w brain MRIs scanned in hospitals of the Greater Paris area (Assistance Publique-Hôpitaux de Paris). We used the same dataset as in our previous study, where we randomly selected 5500 images that were acquired on various scanners (Siemens, GE, Philips and Toshiba).[4] Motion artefacts were manually annotated as a three-grade level by two annotators. A score of (0) was given when no motion was seen, (1) when the structures of the brain were distinguishable despite the presence of motion and (2) when the structures were difficult to distinguish. Some of the 5500 images did not correspond to T1w brain MRI and were therefore not labelled with a motion score (straight reject).[4] If the users labelled differently a given MRI, the consensus grade was chosen as the maximum of the two grades. Dataset characteristics are shown in Table 1.

Table 1. Distribution of the sex and age over the research and clinical datasets. We grouped the three research datasets into a single row.

Database	N patients	N images	Age in years [range]	Sex (%F)
Research databases	136	1222	72.52 [24, 90]	42.57%
Clinical data warehouse	4177	5500	55.51 [18, 95]	55.39%

2.2 Image preprocessing

The T1w MR images were pre-processed using Clinica [14] and its `t1-linear` pipeline. First, a bias field correction was applied using the N4ITK method.[15] An affine registration to the MNI space was then performed.[16] Next, images were cropped to remove background resulting in images of size 169×208×179, with 1 mm isotropic

voxels.[17] The Z-score method, which consists of subtracting the mean intensity of the entire image from each voxel value and dividing it by the corresponding standard deviation, was used to normalise the voxel intensities.

2.3 Proposed approach

We developed a transfer learning approach to detect motion artefacts in clinical images based on motion simulated on research images.

Head motion can be well approximated as rigid body motion which requires six degrees of freedom, comprising three translations and three rotations.[10] We used the Python library TorchIO and its `RandomMotion` function.[18] In this image-based approach, the assumption was made that the subject takes Nt different positions during the acquisition.[8] The motion simulation algorithm follows these four steps:

1. Nt rigid transformations of the motion free image were computed in the image domain. The three translation (t_x, t_y, t_z) and rotation $(\Theta_x, \Theta_y, \Theta_z)$ parameters were sampled for each transformation from a uniform distribution across the translation and rotation ranges given as input.
2. For the Nt transformed images and the original one, the fast Fourier transform was estimated to transform the image domain into the k-space domain.
3. A new k-space was built by concatenating blocks for the Nt different simulated positions.
4. To transform the k-space to the image domain, an inverse fast Fourier transform was calculated to produce the final motion corrupted synthetic image.

To classify motion artefacts, we implemented a CNN composed of five convolutional blocks and three fully connected layers (denoted as Conv5FC3) that proved successful in our previous work.[4] Each convolutional block is made of a convolutional layer, a batch normalisation layer, a ReLU activation function and a max pooling layer. We used the ADAM optimiser, the weighted binary cross-entropy loss and a batch size of 16. Our implementation was done using Pytorch and the ClinicaDL software.[19]

Transfer learning was used to generalise our model from synthetic to routine clinical data. We fine-tuned our pre-trained model by allowing the training of the three fully connected layers of the Conv5FC3 model and freezing the weights of the other layers.

2.4 Experiments

At first, we aimed to test the ability of the proposed Conv5FC3 network to detect motion in research datasets using only synthetic motion while training. We performed a series of experiments on the research datasets, where we corrupted motion-free MRIs with different severities of motion to study the influence of the translation and rotation ranges. A first independent test set was created by randomly selecting 182 images over the three public datasets and corrupting 91 MRIs with different severities of motion (rotation: [2°, 8°]; translation: [2 mm, 8 mm])). The remaining images were split into training and validation using a 5-fold cross validation. Our model was also validated on a second small test set with 41 ADNI MRIs corrupted by real motion.

Then, a second set of experiments was performed on the CDW dataset using transfer learning. We fine-tuned our pre-trained model using 5000 clinical images on two target tasks: the detection of severe (Motion01vs2) and mild (Motion0vs1) motion. We tested our method on an independent test set of 500 clinical MRIs.

3. RESULTS

3.1 Validation on research dataset

We evaluated the performance of our model trained on synthetic motion on our independent test set corrupted with different severities of motion artefacts. We studied the influence of the `RandomMotion` function parameters by performing several experiments with different motion severities. We first trained a model with synthetic severe motion by applying a large rotation and translation range ([6°,8°]; [6 mm,8 mm]). The balanced accuracy (BA) on our independent test set is excellent (98%). We also obtained very good results for smaller ranges of rotation and translation as reported in Table 2. Then, we evaluated the ability of these models to detect real motion. As mentioned in Section 2.1, we defined a test set according to the IPMOTION. Our models were perfectly able to detect motion on these images. In light of these experiments, we concluded that the most suitable model to continue our study was the one trained with rotation and translation ranges of [6°, 8°] and [6 mm, 8 mm], respectively.

Table 2. Results for the detection of synthetic and real motion in T1w brain MRI from research datasets. For the validation on synthetic motion, we report the mean and the empirical standard deviation across the five folds for the balanced accuracy (BA), specificity and sensitivity. For the detection of real motion, only the accuracy obtained by the best model of the 5-fold CV was reported as our independent test set contained only images with motion.

		Cross-validation on synthetic motion			**Test on real motion**
Rotation range	Translation range	BA	Specificity	Sensitivity	Accuracy
(6°, 8°)	(6 mm, 8 mm)	98.22 ± 1.39	99.29 ± 0.86	97.14 ± 2.33	100%
(4°, 6°)	(4 mm, 6 mm)	97.06 ± 1.47	98.25 ± 1.92	95.87 ± 1.27	100%
(2°, 4°)	(2 mm, 4 mm)	95.51 ± 2.47	98.94 ± 2.11	92.06 ± 5.76	98.41%

3.2 Application to the clinical data warehouse

The results obtained with the proposed transfer learning framework on our independent clinical test set are presented in Table 3. For the detection of severe motion, the classifier BA is almost as good as that of the annotators, which is defined as the average of the BA between each rater and the consensus. For the detection of mild motion the classifier BA is low (62.61%) and lower than that of the raters (73.21%). We compared the performance obtained with and without fine tuning. When applying the network trained on synthetic data directly on the clinical data, we observed a large drop in performance with a particularly low specificity for both tasks. A second comparison was performed between the proposed transfer learning framework and when training with the clinical data from scratch. Our transfer learning method achieved a gain of more than 10 percent points for the detection of severe motion. A much smaller improvement was observed for the detection of mild motion.

Table 3. Detection of motion artefacts on the CDW. For both the detection of severe motion (Motion01vs2) and mild motion (Motion0vs1), we report: the agreement between human raters and the consensus (manual annotations), results of the proposed approach (pre-training on synthetic motion from research data and fine tuning on CDW), results when training on synthetic motion from research datasets without fine tuning and results when training from scratch on CDW.

		BA	Specificity	Sensitivity
Motion01vs2	Manual annotations	86.29%	–	–
	Fine-tuning on CDW (proposed)	84.52%	85.37%	83.67%
	Training on research dataset	60.26%	33.33%	87.19%
	Training from scratch on CDW	73.75%	49.58 %	97.93%
Motion0vs1	Manual annotations	73.21%	–	–
	Fine-tuning on CDW (proposed)	62.61%	52.00%	73.23%
	Training on research dataset	53.18%	17.96%	88.57%
	Training from scratch on CDW	58.31%	33.39%	83.24%

4. DISCUSSION AND CONCLUSION

In this work, we developed a transfer learning framework for the automatic detection of motion in 3D T1w brain MRI using a CDW. After having pre-trained a CNN to detect motion artefacts on images from research datasets corrupted with synthetic motion, we improved the generalisation ability of our network on a clinical dataset of 5000 images using fine-tuning. We validated the proposed approach by applying it on 500 manually labelled clinical MRIs. To the best of our knowledge, we are the first to propose a very large-scale validation using a CDW for motion artefact detection using synthetic motion. For the synthetic motion detection on the research datasets, our model achieved excellent results with a BA of 98.22%. Trained using synthetic motion only, the model had no difficulty generalising to real motion on a small research test set (BA: 100%) but it was not able to generalise to the CDW (BA: 60.26%). This poor result with a low specificity highlights both the critical importance of validating models trained on research datasets to clinical ones, but also the quality gap between research, where strict acquisition protocols are respected, and clinical data, which suffer from a lack of homogenisation of acquisition parameters. To overcome these issues, we proposed a transfer learning framework which achieved very good results for the detection of severe motion with a BA of 84.52%, which is nearly as good as that of the annotators (86.29%) and 10 percent points higher than when training the model from scratch.

ACKNOWLEDGMENTS

The research was done using the Clinical Data Warehouse of the Greater Paris University Hospitals. The authors are grateful to the members of the AP-HP WIND and URC teams, and in particular Stéphane Bréant, Florence Tubach, Jacques Ropers, Antoine Rozès, Camille Nevoret, Christel Daniel, Martin Hilka, Yannick Jacob, Julien Dubiel and Cyrina Saussol. They would also like to thank the "Collégiale de Radiologie of AP-HP" as well as, more generally, all the radiology departments from AP-HP hospitals.
The research leading to these results has received funding from the Abeona Foundation (project Brain@Scale), from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6).

APPRIMAGE Study Group

Olivier Colliot, Ninon Burgos, Simona Bottani, Sophie Loizillon [1]
Didier Dormont [1,2], Samia Si Smail Belkacem, Sebastian Ströer [2]
Nathalie Boddaert [3]
Farida Benoudiba, Ghaida Nasser, Claire Ancelet, Laurent Spelle [4]
Hubert Ducou-Le-Pointe[5]
Catherine Adamsbaum[6]
Marianne Alison[7]
Emmanuel Houdart[8]
Robert Carlier [9,17]
Myriam Edjlali[9]
Betty Marro[10,11]
Lionel Arrive[10]
Alain Luciani[12]
Antoine Khalil[13]
Elisabeth Dion[14]
Laurence Rocher[15]
Pierre-Yves Brillet[16]
Paul Legmann, Jean-Luc Drape [18]
Aurélien Maire, Stéphane Bréant, Christel Daniel, Martin Hilka, Yannick Jacob, Julien Dubiel, Cyrina Saussol, Rafael Gozlan [19]
Florence Tubach, Jacques Ropers, Antoine Rozès, Camille Nevoret [20]

[1] Sorbonne Université, Institut du Cerveau - Paris Brain Institute, Inserm, CNRS, AP-HP, Hôpital de la Pitié Salpêtrière, Inria, Aramis project-team, F-75013, Paris, France
[2] AP-HP, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France
[3] AP-HP, Hôpital Necker, Department of Radiology, F-75015, Paris, France
[4] AP-HP, Hôpital Bicêtre, Department of Radiology, F-94270, Le Kremlin-Bicêtre, France
[5] AP-HP, Hôpital Armand-Trousseau, Department of Radiology, F-75012, Paris, France
[6] AP-HP, Hôpital Bicêtre, Department of Pediatric Radiology, F-94270, Le Kremlin-Bicêtre, France
[7] AP-HP, Hôpital Robert-Debré, Department of Radiology, F-75019, Paris, France
[8] AP-HP, Hôpital Lariboisière , Department of Neuroradiology, F-75010, Paris, France
[9] AP-HP, Hôpital Raymond-Poincaré, Department of Radiology, F-92380, Garches, France
[10] AP-HP, Hôpital Saint-Antoine, Department of Radiology, F-75012, Paris, France
[11] AP-HP, Hôpital Tenon, Department of Radiology, F-75020, Paris, France
[12] AP-HP, Hôpital Henri-Mondor, Department of Radiology, F-94000, Créteil, France
[13] AP-HP, Hôpital Bichat, Department of Radiology, F-75018, Paris, France
[14] AP-HP, Hôpital Hôtel-Dieu, Department of Radiology, F-75004, Paris, France
[15] AP-HP, Hôpital Antoine-Béclère, Department of Radiology, F-92140, Clamart, France
[16] AP-HP, Hôpital Avicenne, Department of Radiology, F-93000, Bobigny, France

[17] AP-HP, Hôpital Ambroise Paré, Department of Radiology, F-92100 104, Boulogne-Billancourt, France
[18] AP-HP, Hôpital Cochin, Department of Radiology, F-75014, Paris, France
[19] AP-HP, WIND department, F-75012, Paris, France
[20] AP-HP, Unité de Recherche Clinique, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France

REFERENCES

[1] Nordlinger, B., Villani, C., and Rus, D., eds., [*Healthcare and Artificial Intelligence*], Springer International Publishing, Cham (2020).

[2] Jannot, A.-S., Zapletal, E., Avillach, P., Mamzer, M.-F., Burgun, A., and Degoulet, P., "The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience," *International Journal of Medical Informatics* **102**, 21–28 (2017).

[3] Mia, M. R., Hoque, A. S. M. L., Khan, S. I., and Ahamed, S. I., "A privacy-preserving National Clinical Data Warehouse: Architecture and analysis," *Smart Health* **23**, 100238 (2022).

[4] Bottani, S., Burgos, N., Maire, A., Wild, A., Strer, S., Dormont, D., and Colliot, O., "Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse," *Medical Image Analysis* **75**, 102219 (2021).

[5] Wood, M. L. and Henkelman, R. M., "Truncation Artifacts in Magnetic Resonance Imaging," *Magnetic resonance in medicine* **2**(6), 517–526 (1985).

[6] Andre, J. B., Bresnahan, B. W., Mossa-Basha, M., Hoff, M. N., Smith, C. P., Anzai, Y., and Cohen, W. A., "Toward Quantifying the Prevalence, Severity, and Cost Associated With Patient Motion During Clinical MR Examinations," *Journal of the American College of Radiology* **12**(7), 689–695 (2015).

[7] Mohebbian, M., Walia, E., Habibullah, M., Stapleton, S., and Wahid, K. A., "Classifying MRI motion severity using a stacked ensemble approach," *Magnetic Resonance Imaging* **75**, 107–115 (2021).

[8] Shaw, R., Sudre, C., Ourselin, S., and Cardoso, M. J., "MRI k-Space Motion Artefact Augmentation: Model Robustness and Task-Specific Uncertainty," in [*Medical Imaging with Deep Learning - MIDL 2018*], (2018).

[9] Oksuz, I., "Brain MRI artefact detection and correction using convolutional neural networks," *Computer Methods and Programs in Biomedicine* **199**, 105909 (2021).

[10] Duffy, B. A., Zhang, W., Tang, H., and Zhao, L., "Retrospective correction of motion artifact affected structural MRI images using deep learning of simulated motion," in [*Medical Imaging with Deep Learning - MIDL 2018*], (2018).

[11] Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jack, C. R., Jagust, W. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., and Weiner, M. W., "Alzheimer's Disease Neuroimaging Initiative (ADNI)," *Neurology* **74**(3), 201–209 (2010).

[12] Commowick, O., Cervenansky, F., and Ameli, R., eds., [*MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*] (2016).

[13] Mercier, L., Del Maestro, R. F., Petrecca, K., Araujo, D., Haegelen, C., and Collins, D. L., "Online database of clinical MR and ultrasound images of brain tumors," *Medical Physics* **39**(6Part1), 3253–3261 (2012).

[14] Routier, A., Burgos, N., Díaz, M., Bacci, M., Bottani, S., El-Rifai, O., Fontanella, S., Gori, P., Guillon, J., Guyot, A., Hassanaly, R., Jacquemont, T., Lu, P., Marcoux, A., Moreau, T., Samper-González, J., Teichmann, M., Thibeau-Sutre, E., Vaillant, G., Wen, J., Wild, A., Habert, M.-O., Durrleman, S., and Colliot, O., "Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies," *Frontiers in Neuroinformatics* **15**, 689675 (2021).

[15] Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C., "N4ITK: Improved N3 Bias Correction," *IEEE Transactions on Medical Imaging* **29**(6), 1310–1320 (2010).

[16] Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C., "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis* **12**(1), 26–41 (2008).

[17] Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., and Colliot, O., "Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation," *Medical Image Analysis* **63**, 101694 (2020).

[18] Pérez-García, F., Sparks, R., and Ourselin, S., "TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine* **208**, 106236 (2021).

[19] Thibeau-Sutre, E., Díaz, M., Hassanaly, R., Routier, A., Dormont, D., Colliot, O., and Burgos, N., "ClinicaDL: An open-source deep learning software for reproducible neuroimaging processing," *Computer Methods and Programs in Biomedicine* **220**, 106818 (2022).