Multicenter Imaging Studies: Automated Approach to Evaluating Data Variability and the Role of Outliers

Mariana Bento, Roberto Souza, Richard Frayne
Radiology and Clinical Neuroscience, Hotchkiss Brain Institute
University of Calgary, Calgary, Canada
Email: {mariana.pinheirobent, roberto.medeirosdeso, rfrayne}@ucalgary.ca

***Abstract*—Magnetic resonance (MR) as well as other imaging modalities have been used in a large number of clinical and research studies for the analysis and quantification of important structures and the detection of abnormalities. In this context, machine learning is playing an increasingly important role in the development of automated tools for aiding in image quantification, patient diagnosis and follow-up. Normally, these techniques require large, heterogeneous datasets to provide accurate and generalizable results. Large, multi-center studies, for example, can provide such data. Images acquired at different centers, however, can present varying characteristics due to differences in acquisition parameters, site procedures and scanners configuration. While variability in the dataset is required to develop robust, generalizable studies (*i.e.*, independent of the acquisition parameters or center), like all studies there is also a need to ensure overall data quality by prospectively identifying and removing poor-quality data samples that should not be included, *e.g.*, "outliers". We wish to keep image samples that are representative of the underlying population (so called "inliers"), yet removing those samples that are not. We propose a framework to analyze data variability and identify samples that should be removed in order to have more representative, reliable and robust datasets. Our example case study is based on a public dataset containing T1-weighted volumetric head images data acquired at six different centers, using three different scanner vendors and at two commonly used magnetic fields strengths. We propose an algorithm for assessing data robustness and finding the optimal data for study occlusion (*i.e.*, the data size that presents with lowest variability while maintaining generalizability (*i.e.*, using samples from all sites)).**

I. Introduction

Machine learning techniques (ML) are becoming increasingly applied to medical imaging. Commonly, they seek to improve patient diagnosis and treatment strategies. [1] Magnetic resonance (MR) imaging, for example, is a commonly used imaging modality that produces images with relatively high spatial resolution and, often, with high image contrast between normal and abnormal tissues. MR is a frequent target of ML processing techniques. [2], [3]

A main issue when developing ML techniques for human imaging data is the inadequately small number of samples available during the training phase. This observation is especially true for deep learning techniques. Even though techniques have been developed in an attempt to overcome this issue, such as data augmentation, [4] if more data were available, then the proposed automatic ML techniques are likely to have improved accuracy and reliability, and generalizability.

Another common issue with ML-based techniques is that they are often developed using fairly homogeneous datasets. In the case of MR-based data, this might arise from using images acquired with specific, tightly controlled parameters, from a single scanner (*i.e.*, vendor, magnetic field strength) at a single site (*i.e.*, single center data). Most recent and more effective studies include different scanners and/or sites in their datasets to assure method robustness and reliability to ensure that the findings are generalizable to different datasets. [5], [6]

The most generalizable solution occurs when combining as much data as possible to best represent the underlying population. However, when including data acquired from multiple centers with possibly varying acquisition parameters, we now face increased problems related to data variability and other quality-control issues. [7] Data acquired at different centers, using different scanner vendors and acquisition parameters, *etc*. tend to have different characteristics, such as varying spatial resolution, image contrast, signal-to-noise ratio, among others. [8] Related studies suggest that these and other factors increase imaging heterogeneity on automatic ML techniques, quantitative measurements and biomarkers. Automated quality control methods are needed to assess data reliability, reproducibility and robustness , [9]–[11] particularly in multicenter studies. [7], [12]–[15]

Pre-processing techniques are commonly applied to overcome some of the larger unwanted sources of variability in the dataset. This key step has the goal of making the data appear as homogeneous as possible, by changing the intensity range (*normalization techniques*), [16] contrast variation (*non-uniformity correction*), [17] fixing a standard size for all image (*image resize*), *etc*. [18], [19]. However, pre-processing techniques do not correct for a more fundamental problem: poor quality samples, commonly revered to as *outliers*.

Including outliers in automated ML techniques, may lead to inaccurate results, and thus support incomplete or even incorrect conclusions. While it is good practice to use large, heterogeneous datasets to develop robust ML techniques, it is most likely that some samples in these datasets should be detected as non representative samples (labeled as outliers) and be removed.

In this work, a framework is proposed to analyze a multi-center imaging dataset, to better understand data variability (due to difference in scanner vendor, field strength and acquisition parameters), and to study what variability is appropriate or inappropriate (by removing data outliers). Similar studies in handwriting classification examined the benefit of removing anomalies in large scale datasets and the risk of decreasing performance due to removing too many samples. [20]

There are also related works in the medical imaging area, that use outlier detection or anomaly detection methods to identify and segment abnormalities within structures of interest. [21], [22] Other methods use outlier detection as a quality control technique and protocol compliance. [23], [24] However, in the best of our knowledge, these is no work in the literature that uses outlier detection methods to study data variability in multicenter imaging MR studies, and to identify prospective outliers.

In our proposed framework, we wish to detect and remove outliers without discarding important and representative image data. Our approach differs from applying pre-processing techniques to images as these methods aim to decrease only inter-sample data variability, must be carefully selected and are application specific. Our approach aims to better understand the sources of variability and detect possible outliers that could cause errors. It seeks to determine the subset of data samples that best represents the underlying population (*i.e.*, healthy control subjects, or a confirmed patient group).

In this proceeding, we present a case study using MR brain images acquired at different sites, using three scanner vendors and at two magnetic field strengths. In addition to studying and analyzing data heterogeneity, we propose a method to find data subsamples that have lower variability (than the original dataset), yet are representative of the underlying population and can selectively presents samples from one to all sites. However, our framework may be used in different applications to better understand the data variability and representativeness.

II. Multicenter dataset

We used a public multi-center dataset comprising volumetric T1-weighted MR images acquired from healthy control subjects (*Calgary-Campinas-359* or *CC-359*, http://miclab.fee.unicamp.br/tools, described in [25]). These data were acquired at six different sites, using scanners from the three common MR equipment vendors (General Electric, Philips, and Siemens, subsequently anonymously labeled as $SV_{1,2,3}$) and at two different field strengths (1.5 T and 3 T) were used (Figure 1). The dataset is composed of images from 359 healthy adults (ranging in age from 29 to 80 years), with approximately 60 subjects per vendor and magnetic field strength combination (Table I). Healthy control subjects were chosen, for this case study, to ensure that our findings were related to image quality and variability differences in acquisition parameters, and not due to brain abnormalities or pathology. In a future and more comprehensive study, we would aim to include in our experimental dataset pathological samples.

TABLE I
Dataset description. Summary of average age (mean ± standard deviation), gender balance (number of male/number of female subjects) and total number of subjects) by scanner vendor, magnetic field strength. [25]

Scanner Vendor	Field Strength	Age	Gender	Subjects
SV_1	1.5 T	53.9 ± 7.3	30M/30F	60
	3.0 T	56.6 ± 6.9	30M/30F	60
SV_2	1.5 T	52.8 ± 9.6	26M/33F	59
	3.0 T	50.0 ± 9.3	30M/30F	60
SV_3	1.5 T	53.9 ± 5.8	30M/30F	60
	3.0 T	53.6 ± 5.7	30M/30F	60
All	1.5 T and 3 T	53.5 ± 7.8	176M/183F	359

III. Analyzing data variability

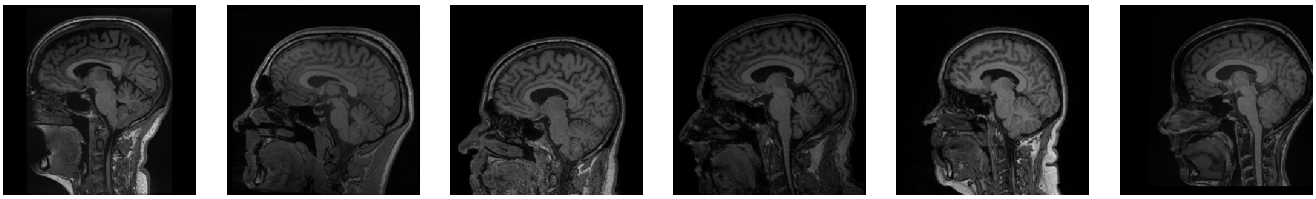
We propose a generic framework to study image variability and to prospectively identify sample outliers that should be removed from the full dataset in order that it have a better description of the underlying population. Our proposal has three main steps: A) pre-processing, B) feature extraction, and C) outlier detection (Figure 2). These steps were recursively performed using a high dimensional convolutional feature space to detect outliers and study data variability.

An important concept we term as "representativity" must be first defined. In our case study, we have six different sites covering all permutations of scanner vendor and magnetic field strength. Our definition of representativity is to have samples from all sites, so that all combinations of acquisition parameters are included in the model. For other applications and/or when using other datasets, it would be possible to choose a different definition of representativity, such as desiring to remove entire sites with known or suspected acquisition problems, susceptibility to motion artifact, or reconstruction issues. By using appropriate definitions of representativity, our proposed framework may be applied to several applications to better understand data variability, and to recognize samples with characteristics that are distinct from the other samples in the dataset. These excluded samples are commonly called outliers, while the remaining included samples are called inliers. [26]

A. Pre-processing and feature extraction

The extraction of features from images is an important task in our proposed framework. This step needs to provide comprehensive and discriminative information about the imaging characteristics, [27] and have good capability for generalization. The features automatically detected in convolutional networks, *i.e.*, convolutional features are generally appropriate [28] and recently have been applied to multicenter dataset studies. [1], [4], [27]

In order to compute the convolutional features, images were first pre-processed using two straight-forward and common applied steps: 1) resizing and 2) intensity normalization. The images were first resized to 224×224 using a bivariate spline interpolation of the first order. [19] Each image then was normalized so that its intensity values lay in the range [0,1].



(a) SV_1; 1.5 T (b) SV_1; 3 T (c) SV_2; 1.5 T (d) SV_2; 3 T (e) SV_3; 1.5 T (f) SV_3; 3 T

Fig. 1. Representative dataset samples from each one of the six vendors and magnetic field combinations.

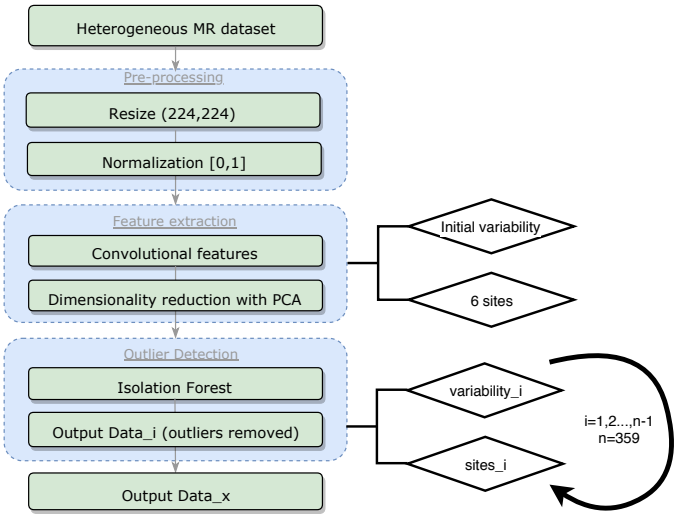


Fig. 2. Flow chart of proposed framework to detect outliers and determine the optimum number of samples that decrease variability yet maintaining representativity (in our example study, including at least one sample from all sites). The step (i) to find and remove single outliers is repeated $n-1$ times ($n = 359$), with each iteration computing the variability ($variability_i$) and the number of sites ($sites_i$) represented in the retained data. The optimal number, in our study, x, represents the i value for which $variability_i$ is reduced and $sites_i = 6$.

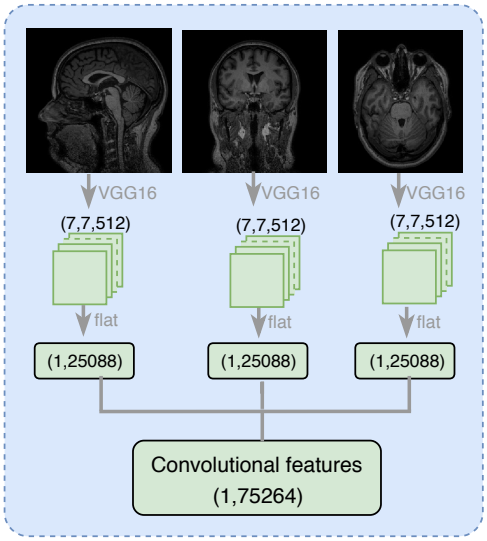


Fig. 3. Convolutional features computed form three orthogonal views (*i.e.*, 2D mid-volume images) per brain volume. The VGG16 network [29] was used to extract convolutional features. For each input slice (of size 224×224), feature maps (of size $7 \times 7 \times 512$) were computed using the network (with the last fully connected layer removed), flatted into an array with $25,088$ elements. Features from each orthogonal view were concatenated to generate the convolutional features array (of $75,264$ features per MR image).

The convolutional features were computed from the images (*i.e.*, 2D MR slices) by using the deep convolutional network (VGG16 [29]), pre-trained with imagenet weights from Ref [4]. This network was used to perform feature extraction by removing the last layer (*i.e.*, the fully connected layer).

In order to have features that characterize the three-dimensional image volume using a 2D approach, we used orthogonal slices. For each image volume, convolutional features were computed for the three orthogonal 2D mid-volume slices (*i.e.*, axial, sagittal and coronal views). The network outputs a feature matrix of size $7 \times 7 \times 512$ for each one of these slice. this matrix as flattened to a one-dimensional feature array of $25,088$ elements. Because three orthogonal slices were selected to describe each volumetric image, a total of $3 \times 25,088 = 75,264$ convolutional features were extracted per volume (Figure 3). This feature space was used to perform the outlier detection method.

We also performed feature dimensionality reduction on the original convolutional feature space to allow feature space visualization and variability measurements. The original feature space ($75,264$ convolutional features) was reduced to two features by using principal component analysis algorithm (PCA) [30]. The resulting 2D plane was defined by the two eigenvectors corresponding to the two largest eigenvalues (*i.e.* the two most relevant principal components, PCA_1 and PCA_2).

B. Outliers detection algorithm

We selected the isolation forest algorithm [31] to automatically detect outliers. This method is based on random forests and is, thus, suitable for large-dimensional settings, such as the proposed feature space with $75,264$ features. The algorithm also was selected because it has demonstrated good results in other applications, [26] but mostly it was used because an input parameter is the percentage (%) of prospective outliers to be removed (allowing the full assessment of the dataset variability).

The algorithm was recursively repeated $i = n - 1$ times by using the original feature space (before PCA feature dimensionality reduction), where n equals the number of samples

in the dataset to be analyzed (Figure 2). Initially $i = 1$. In each successive iteration, the next most outlying sample was labeled as outlier and removed. The algorithm was repeated until there was only one sample left in the data (last iteration $i = n - 1$; note in our case study $n = 359$).

In each iteration, the most outlying sample was removed, and two measurements were re-computed on the remaining data: 1) the number of sites that were still represented ($sites_i$) representativity) and 2) the variance of ($variability_i$, measurement of data dispersion). The variance (Var) was defined as the standard deviation of the remaining samples in the two dimensional PCA feature space:

$$Var = \sqrt{\frac{1}{(n-i)} \sum_{i=1}^{n-1} ((r_{1,i} - \bar{r_1})^2 + (r_{2,i} - \bar{r_2})^2)} \quad (1)$$

where n defines the total number of samples. Data in the two-dimensional feature space (PCA_1 and PCA_2) is represented by $r_{1,i}$ and $r_{2,i}$ respectively, and $\bar{r_1}$ and $\bar{r_2}$ are the mean values of PCA_1 and PCA_2. At each iteration, $\bar{r_1}$ and $\bar{r_2}$ are recomputed in the remaining samples.

After completing the algorithm, data variability (Var) was normalized to lie between zero and one to facilitating visualization and comparisons. It was expected that the variability curve would be a decreasing curve, since outliers were being removed, thus the remaining data should be more homogeneous samples (*i.e.*, consisting of inliers).

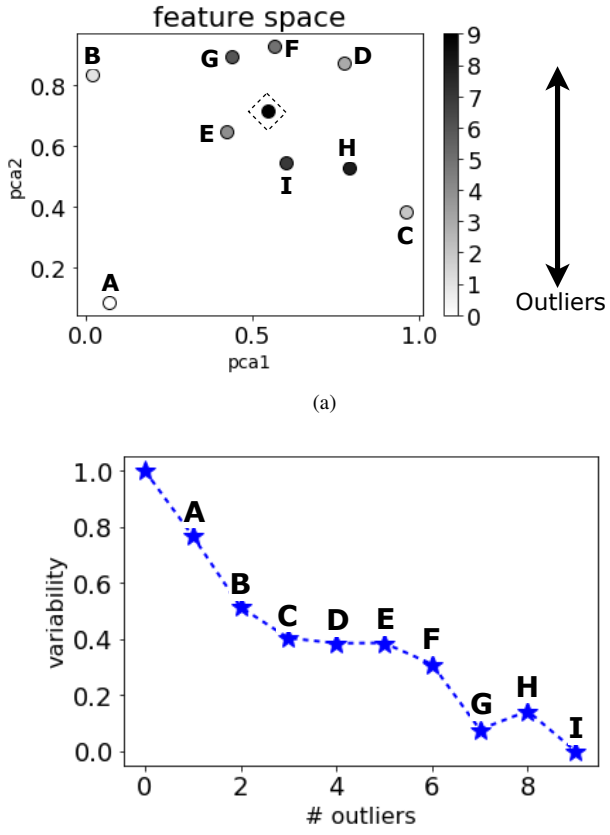
This outlier detection algorithm was first run on simulated data (a randomly generated dataset containing 10 samples) and then on the *CC-359* dataset.

IV. Results

While our proposed framework is generalizable to other applications, our results and discussion were based on a synthetic study and one case study using a multi-vendor and multi-magnetic field dataset. Our goal in both studies was to understand dataset variability, and to find an optimal sample size with lowest variability. In the multi-vendor study, we set the additional criterion of maintaining samples from all sites (lowest variability and preserving representativity).

A. Synthetic dataset

In an initial study using synthetic data, the decreasing variability curve and the corresponding feature space are presented in Figure 4. The first sample detected as an outlier, identified as point A, is geometrically the most distant sample in the feature space (as expected). The variability computed after removing A is presented in Figure 4b. After each iteration ($i = 1, \ldots 9$) in the recursive algorithm, another sample was removed, and the variability computed for the remaining samples. In the last iteration, sample I is extracted, and a single sample is left (dashed diamond in Figure 4a). The variability, by definition, now equals zero.



(a)

(b)

Fig. 4. Synthetic data example. Ten (10) samples (randomly generated) to illustrate the outlier detection method by showing (a) the correspondence in the feature space and (b) the variability computed at the end of each iteration via the variability *versus* outliers removed curve. Letters (A through I) were used to identify samples and the correspondences between figures. The dashed diamond in (a) identifies the only sample left after the final iteration.

B. CC-359 dataset

Using the *CC-359* data, the outlier detection algorithm was applied. After each iteration, the variance and the number of sites still represented in the remaining data was recorded. After some iterations (67 in this case study), data for a site were completely removed (Table II).

TABLE II
NUMBER OF REMOVED OUTLIERS AND THE NUMBER OF REMAINING SITES REPRESENTED BY THE INLIERS. FOR EXAMPLE, WHEN THE NUMBER OF OUTLIERS WAS BETWEEN 1 AND 67, ALL SIX SITES WERE PRESENTS; BETWEEN 68-120, ONLY FIVE SITES WERE PRESENTED, *etc*.

# Outliers	Remaining Sites
1-67	Six
68-120	Five
121-332	Four
333-350	Three
351-357	Two
358	One

Samples from two sites were completely removed (120 samples), before any samples from the other four sites were

removed. This observation provides relevance on the variability in this multicenter dataset. In practice, it would warrant further investigation of the two poorly performing sites.

In our case study, it was necessary to remove more than two hundred samples to completely discard a third site (there are four sites represented in the samples between algorithms iterations 121 and 332). From this we can conclude that these four sites are more homogeneous.

Besides analyzing the number of sites still represented after each iteration, the normalized variability of the remaining data was also evaluated *via* the two dimensional standard deviation of the PCA space, (Figure 5). As expected the variability was found to a nearly always decreasing curve.

Because in our case of study, we desire to maintain samples from all sites (fulfill the representativity requirement), the optimum number of removed outliers was 67. This number of discarded samples gave the lowest variability, while still keeping samples from all sites (Figure 5). Lower variabilities were possible but did not include samples from all six sites.

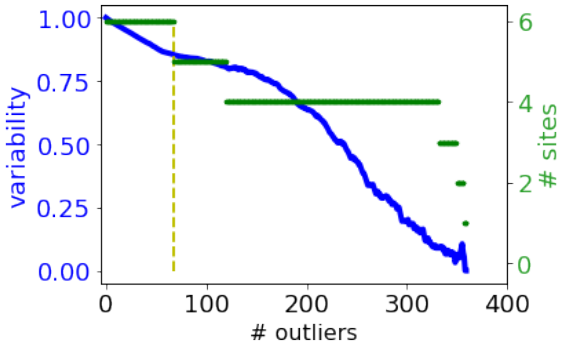
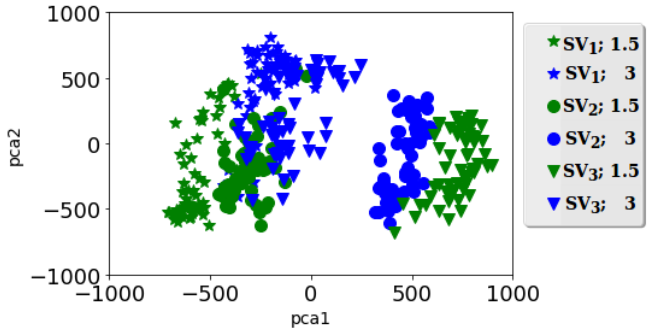


Fig. 5. Computed data variability (blue line) and the number of represented sites (green line) when removing outliers. As expected, when removing outliers, the variability decreases because the remaining dataset becomes more homogeneous. However, in the early iterations of this case study, all samples for a site were eliminated, decreasing the representativity of the experimental dataset. Our goal was to find optimum number of removed outliers that gives minimum variability, while still presenting samples from all sites (yellow vertical dashed line)

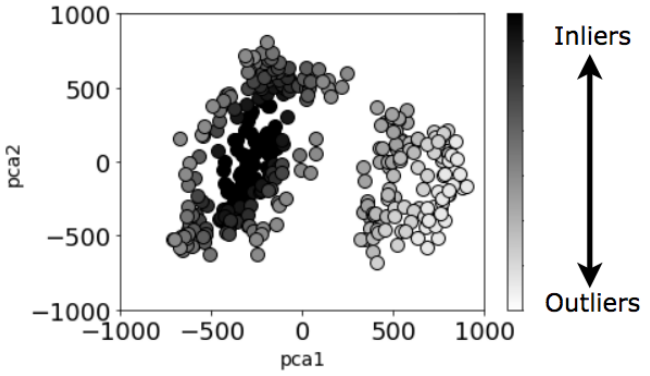
A more qualitative analysis of the operation of our outlier detection algorithm is possible by visualizing the two dimensional feature space computed by using PCA to reduce the convolutional features dimensionality (Figure 6). The original feature space shows the original dataset heterogeneity (data dispersion). It is possible to visualize that sample from some sites (vendor SV_3; 1.5T and SV_2; 3T) are separated from samples from other sites (Figure 6a). We also present the feature space while removing outliers (Figure 6b). As expected samples that were more easily distinguished from other samples (blue circles and green triangles in Figure 6a) were the first ones to be removed (brighter circles in Figure 6b).

V. Discussion

We have proposed a framework to study data variability, detect possible outliers and give a more homogeneous result suitable for use with heterogeneous and/or multicenter



(a)



(b)

Fig. 6. 2D Feature space visualization: (a) considering the original data with six sites (three vendors: SV_1 to SV_3; two magnetic field strength: 1.5T and 3T), and (b) the order in which outliers were removed using a gray scale encoding from black to white. Brighter circles represent samples that were removed in initial iterations (outliers), and darker samples were maintained until later iterations (inliers).

datasets. While different datasets will, in general, present with different variability curves, we have demonstrated that the algorithm will still work as intended.

Specifically in our case study, we observed a decreasing variability curve until almost all samples were removed. In the last few iterations, the outlier detection method still performed as expected, however some non-monotonic performance was seen in the variability curve because there are now too few samples in the remaining data (resulting in the small spikes seen in later iterations of Figure 5). In our case of study, it was possible to verify that heterogeneous datasets acquired in multiple sites present varying characteristics, with some sites presenting images that were more similar than others (clusters). However, keeping all sites makes the dataset more representative and generalizable allowing a multicenter and multi magnetic field strength study, considering none of the sites were corrupted, or presented systematic quality problems.

Depending on the application (study goals and hypothesis), the definition of representativity and, in consequence, the definition of outlier may change, making this framework data-driven and suitable for other applications.

It is quite possible in some large multi-center imaging studies to have poorly performing sites that should be com-

pletely removed, or otherwise not considered in any analysis. These sites would typically contain images of sufficiently low quality images, images with artifacts, or those that were wrongly acquired or reconstructed, to warrant exclusion. These sites would be easily identified in our framework, because all their samples would be initially removed. In contrast, multicenter dataset containing similar samples (homogeneous dataset) should present a variability curve with a more slowly decreasing shape, and samples from all sites would be maintained until the later iterations.

Further analysis and methods based on the output dataset (after removing outliers) are more likely to produce more robust findings. For example, classification of control subjects from patient is more likely to be based on abnormalities (pathology) than being based on imaging characteristics. This raises the hypothesis that by removing erroneous samples (using the proposed outlier detection method as a preliminary step in a basic ML tool) would lead to more robust and significant results. However, further analysis is necessary to prove this hypothesis.

VI. Conclusions

Our proposed framework analyzes large and multicenter imaging data, studying its variability, identifying prospectively outliers, and outputting a more homogeneous dataset. This output may provide a better representation of specific groups of interest. The outlier detection algorithm improves the combination of data acquired using different acquisition parameters, site and scanner vendor (mostly in a multicenter setting), by proposing a set of samples that present with lower variability yet, if desired, guaranteeing data representativity.

In our case study, we evaluated a combination of data acquired using different scanner vendors and magnetic field strengths. We found a subset of samples with reduced variability, while maintaining samples from all sites (preserving representativity in a multicenter and multi magnetic field strength study). The definition of representativity is important and may change, or be specific for each application.

Future work will analyze larger datasets containing control subjects and patients acquired at many sites. We will investigate other feature and, possibly, combinations of multiple feature to have a more comprehensive description of the data to be used in the variability analysis. Finally, we also intend to evaluate the usage of the proposed outlier detection method as a preliminary step in a machine learning technique, comparing the results when removing or not poor-quality samples, identified as outliers.

Acknowledgment

The Calgary-Campinas collaboration and the *CC-359* dataset was supported in part by an award from Coordination for the Improvement of Higher Education Personnel (CAPES, Brazil), Special Visiting Professor Program (PVE-88881.062158/2014-01). The CNS was funded by the Canadian Institutes for Health Research (CIHR). Mariana Bento, PhD was supported by the Hotchkiss Brain Institute (HBI) and Roberto Souza, PhD has a T. Chen Fong 2018 Postdoctoral Fellowship. Richard Frayne, PhD was supported by the Hopewell Professorship in Brain Imaging. Computational infrastructure at the Calgary Image Processing and Analysis Centre (CIPAC) was supported by the Canadian Foundation for Innovation (CFI) and the Government of Alberta.

References

[1] G. Litjens, T. Kooi, B. Bejnordi, A. Setio, F. Ciompi, M. Ghafoorian, J. van der Laak, B. van Ginneken, and C. Sanchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, no. 1, pp. 60–88, 2017.

[2] M. Wernick, Y. Yang, J. Brankov, G. Yourganov, and S. Strother, "Machine learning in medical imaging," *IEEE signal processing magazine*, vol. 27, no. 4, pp. 25–38, 2010.

[3] M. Leite, L. Rittner, S. Appenzeller, H. Ruocco, and R. Lotufo, "Etiology-based classification of brain white matter hyperintensity on magnetic resonance imaging," *Journal of Medical Imaging*, vol. 2, no. 1, pp. 014 002–1:10, 2015.

[4] H. C. Shin, H. R. Roth, M.Gao, L. Lu, Z. Xu, I.Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[5] G. L., G. Zamboni, A. Khan, L. Li, G. Bonifacio, V. Sundaresan, U. Schulz, W. Kuker, M. Battaglini, P. Rothwell, and M. Jenkinson, "BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities," *Neuroimage*, vol. 141, pp. 191–205, 2016.

[6] M. Bento, R. Souza, R. Lotufo, R. Frayne, and L. Rittner, "Wmh segmentation challenge: A texture-based classification approach," in *Lecture Notes in Computer Science: Proceedings of International MICCAI Brain lesion Workshop*, 2017.

[7] M. Davids, F. Zollner, M. Ruttorf, F. Nees, H. Flor, G. Schumann, and L. Schad, "Fully-automated quality assurance in multi-center studies using MRI phantom measurements," *Magnetic Resonance Imaging*, vol. 32, no. 6, pp. 771–80, 2014.

[8] M. Zelikman, S. Kruchinin, and K. Snopova, "Methodology and tools for quality control of magnetic resonance imaging devices," *Biomedical Engineering*, vol. 44, no. 5, pp. 184–187, 2011.

[9] F. Guio, E. Jouvent, G. Biessels, S. Black, C. Brayne, C. Chen, C. Cordonnier, F. Leeuw, M. Dichgans, F. Doubal, M. Duering, C. Dufouil, E. Duzel, F. Fazekas, V. Hachinski, M. Ikram, J. Linn, P. Matthews, B. Mazoyer, V. Mok, B. Norrving, J. OBrien, L. Pantoni, S. Ropele, P. Sachdev, R. Schmidt, S. Seshadri, E. Smith, L. Sposato, B. Stephan, R. Swartz, C. Tzourio, M. Buchem, A. Lugt, R. Oostenbrugge, M. Vernooij, A. Viswanathan, D. Werring, F. Wollenweber, J. Wardlaw, and H. Chabriat, "Reproducibility and variability of quantitative magnetic resonance imaging markers in cerebral small vessel disease," *Journal of Cerebral Blood Flow & Metabolism*, vol. 36, no. 8, pp. 1319–1337, 2016.

[10] J. Song, S. Hwang, G. Chung, and G. Jin, "Intra-Individual, Inter-Vendor Comparison of Diffusion-Weighted MR Imaging of Upper Abdominal Organs at 3.0 Tesla with an Emphasis on the Value of Normalization with the Spleen," *Korean Journal of Radiology*, vol. 17, no. 2, pp. 209–217, 2016.

[11] R. Heinen, W. Bouvy, A. Mendrik, M. Viergever, G. Biessels, and J. Bresser, "Robustness of Automated Methods for Brain Volume Measurements across Different MRI Field Strengths," *Plos One*, vol. 11, no. 10, p. e0165719, 2016.

[12] K. Helmer, M. Chou, R. Preciado, B. Gimi, R. N., A. Song, J. Turner, and S. Mori, "Multi-site study of diffusion metric variability: effects of site, vendor, field strength, and echo time on regions-of-interest and histogram-bin analyses," in *Proceedings of SPIE–the International Society for Optical Engineering*, 2016.

[13] C. Schlett, T. Hendel, J. Hirsch, S. Weckbach, S. Caspers, J. Menger, T. Ittermann, F. Brenkenhoff, S. Ladd, S. Moebus, C. Stroszczynski, B. Fischer, M. Leitzmann, C. Kuhl, F. Pessler, D. Hartung, Y. Kemmling, H. Hetterich, K. Amunts, M. Gunther, F. Wacker, E. Rummeny, H. Kauczor, M. Forsting, H. Volzke, N. Hosten, M. Reiser, and F. Bamberg, "Quantitative, Organ-Specific Interscanner and Intrascanner Variability for 3 T Whole-Body Magnetic Resonance Imaging in a Multicenter,

Multivendor Study," *Investigative Radiology*, vol. 51, no. 4, pp. 255–265, 2016.

[14] S. Chalavi, A. Simmons, H. Dijkstra, G. Barker, and A. Reinders, "Quantitative and qualitative assessment of structural magnetic resonance imaging data in a two-center study," *BMC Medical Imaging*, vol. 12, no. 27, pp. 1–15, 2012.

[15] P. Tofts and D. Collins, "Multicentre imaging measurements for oncology and in the brain," *The British Journal of Radiology*, vol. 84, pp. 213–226, 2011.

[16] B. Belaroussi, J. Milles, S. Carme, Y. M. Zhu, and H. Benoit-Cattin, "Intensity nonuniformity correction in MRI: Existing methods and their validation," *Medical Image Analysis*, vol. 10, no. 2, pp. 234 – 246, 2006.

[17] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: improved N3 bias correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310 – 1320, 2010.

[18] R. Woods and R. C. Gonzalez, *Digital Image Processing*, P. Hall, Ed. Edgard Blucher, 2000.

[19] S. Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, 2014.

[20] Z. Li, X. Zhang, H. Muller, and S. Zhang, "Large-scale retrieval for medical image analytics: A comprehensive review," *Medical Image Analysis*, vol. 43, pp. 66–84, 2018.

[21] C. Bowles, C. Qin, R. Guerrero, R. Gunn, A. Hammers, D. Dickie, M. Hernndez, J. Wardlaw, and D. Rueckert, "Brain lesion segmentation through image synthesis and outlier detection," *Neuroimage: Clinical*, vol. 16, pp. 643–658, 2017.

[22] K. Li, C. Ye, Z. Yang, A. Carass, S. Ying, and J. Prince, "Quality Assurance using Outlier Detection on an Automatic Segmentation Method for the Cerebellar Peduncles," in *Proceedings of SPIE International Society of Optics and Photonics*, 2016.

[23] M. Niethammer, S. Bouix, S. Fernandez, C. Westin, and M. Shenton, "Outlier Rejection for Diffusion Weighted Imaging," in *Proceedings of Medical Image Computing and Computer-Assisted Intervention*, 2007.

[24] D. Roalf, M. Quarmley, M. Elliott, T. Satterthwaite, S. Vandekar, K. Ruparel, E. Gennatas, M. Calkins, T. Moore, R. Hopson, K. Prabhakaran, C. Jackson, R. Verma, K. Hakonarson, R. Gur, and R. Gur, "The Impact of Quality Assurance Assessment on Diffusion Tensor Imaging Outcomes in a Large-Scale Population-Based Cohort," *Neuroimage*, vol. 125, p. 903919, 2016.

[25] R. Souza, O. Lucena, J. Garrafa, D. Gobbi, M. Saluzzi, S. Appenzeller, L. Rittner, R. Frayne, and R. Lotufo, "An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement," *NeuroImage*, vol. 170, pp. 482–494, 2018.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 1, pp. 2825–2830, 2011.

[27] D. Shen, G. Wu, and H. I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[28] T. Yichuan, "Deep learning using linear support vector machines," in *Proceedings of International Conference on Machine Learning*, 2013.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of 3rd International Conference on Learning Representations*, 2015.

[30] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, p. 301, 1901.

[31] F. Liu, T. Tony, M. Kai, and Z. Zhou, "Isolation forest," in *Proceedings of IEEE International Conference on Data Mining*, 2008.