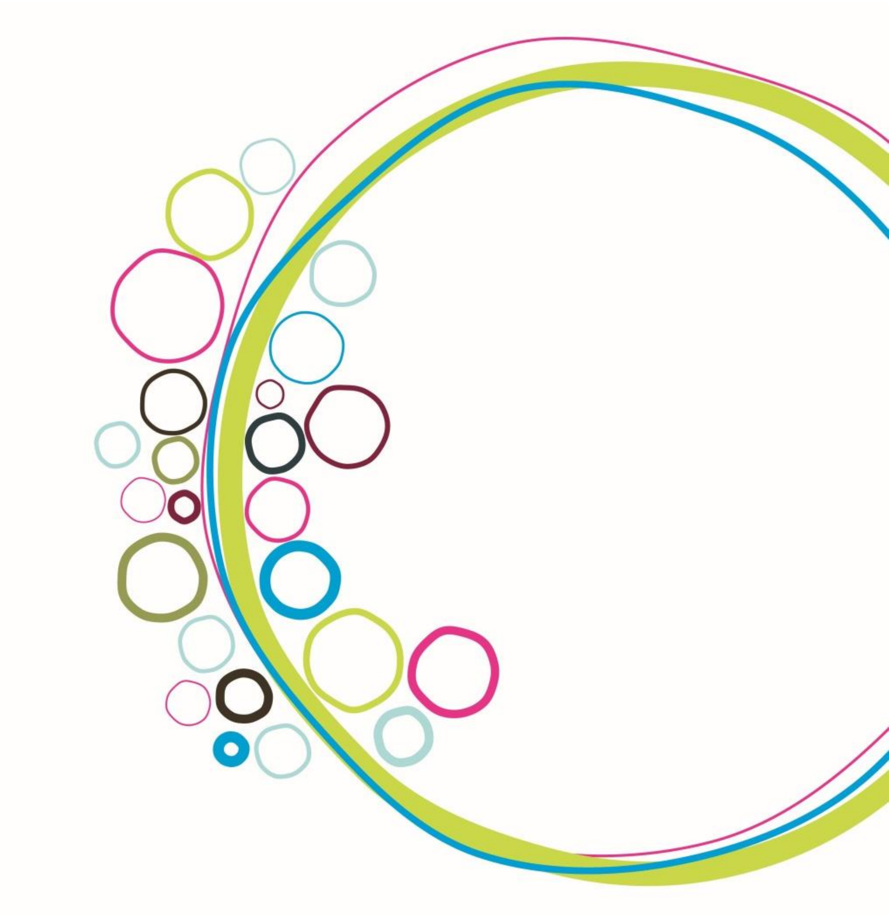


FULL-HD:

FULL EXPLOITATION OF HIGH-DIMENSIONALITY IN BRAIN IMAGING

Report of a JPND Working Group on Harmonisation and Alignment in Brain Imaging Methods

April 2018

JPND
research
EU Joint Programme – Neurodegenerative Disease Research

Working Group title	**Full exploitation of High-Dimensionality in brain imaging**
Acronym	**Full-HD**

FINAL REPORT

High-dimensionality in neurodegenerative disease research

Technological innovations have enabled large-scale acquisition of biological information from human subjects. The emergence of these big datasets has resulted in various ‘omics’ fields. Systematic and large-scale investigations of DNA sequence variations (genomics), gene expression (transcriptomics), proteins (proteomics), small molecule metabolites (metabolomics), and medical images (radiomics), among other data, lie at the basis of many recent biological insights. These analyses are typically unidimensional, i.e. studying one of these ‘omics’ in relation to a neurodegenerative disease. Although this approach has proven its scientific merit through many discoveries, jointly investigating multiple big datasets would allow for their full exploitation, as is increasingly recognized throughout the ‘omics’ world (Medland et al. Nat Neuro. 2014). However, the ultra-high-dimensional nature of these analyses made them challenging and unfeasible in current research settings.

In the HD-READY consortium, i.e. our previous JPND working group, we specifically focused on the computational and statistical requirements for analyzing high-dimensional data, which are far beyond the infrastructural capabilities for single sites. The work performed in HD-READY was extremely successful, resulting in two key publications of novel methods and a software package (“HASE”) that overcome these hurdles (Adams et al. bioRxiv. 2016; Roshchupkin et al. bioRxiv. 2016). For example, associating 1.5 million neuroimaging phenotypes with 9 million genetic variants using the HASE software is now possible in several hours instead of years, with great reductions in the size of data to transfer (Gigabytes instead of Terabytes).

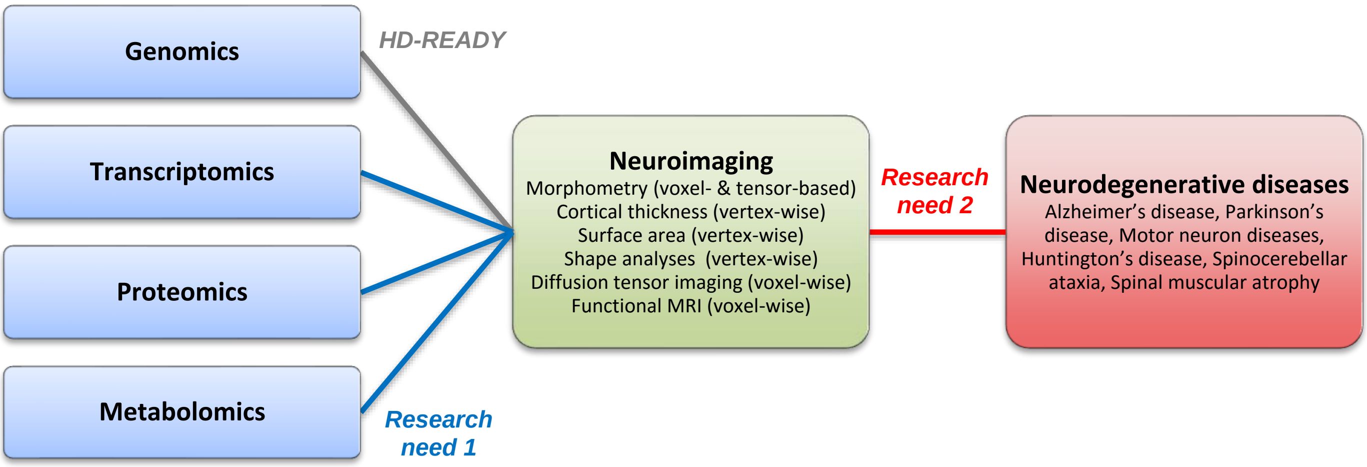
Harmonization of high-dimensional neuroimaging data: an *a priori* approach

Tools delivered in HD-READY are tailor-made to tackle challenges posed by high-dimensional data. However, especially for large-scale imaging datasets, an important outstanding issue is the urgent need to establish a framework for harmonization (Full-HD). Although different data collection and processing complicates comparisons in neuroimaging studies in general, it is of particular importance for high-dimensional data. For example, while gross hippocampal volumes obtained using different methods can still be compared to some extent, it becomes impractical to compare a certain hippocampal voxel with one from another dataset that was processed differently.

Laudable (JPND) efforts such as STRIVE and METACOHORTS that aimed to harmonize vascular imaging markers have typically followed decades of research using heterogeneous methods. Furthermore, the focus of such efforts was purely on aggregate neuroimaging measures, while voxel-wise or vertex-wise harmonization remains elusive. The field of high-dimensional research is relatively young, but growing rapidly, and would greatly benefit from such a harmonization effort early on. Specifically, the Full-HD working group addressed **two research needs**:

1) To harmonize high-dimensional neuroimaging data so that it can be combined with other omics data. Given that a wealth of data has already been acquired using different scanners, field strength, and acquisition protocols, we will set out to define a general framework to harmonize currently available high-dimensional phenotypes but also provide requirements for future/novel neuroimaging phenotypes. Voxel-wise and vertex-wise phenotypes will have a central focus.
2) To harmonize ultra-high-dimensional neuroimaging-by-omics data for neurodegeneration research. We foresee these neuroimaging-by-omics datasets becoming useful tools for neurodegeneration research. For example, if a particular brain atrophy pattern is detected in certain patients, it could be interesting to examine whether there are genetic variants giving rise to a similar pattern. Thus, it is essential that any framework for harmonization of such high-dimensional data should take the ease of use for other researchers into account.

Figure 1 | Role of high-dimensional data in neurodegeneration research.


Genomics
Transcriptomics
Proteomics
Metabolomics
HD-READY
Research need 1
Neuroimaging
Morphometry (voxel- & tensor-based)
Cortical thickness (vertex-wise)
Surface area (vertex-wise)
Shape analyses (vertex-wise)
Diffusion tensor imaging (voxel-wise)
Functional MRI (voxel-wise)
Research need 2
Neurodegenerative diseases
Alzheimer's disease, Parkinson's disease, Motor neuron diseases, Huntington's disease, Spinocerebellar ataxia, Spinal muscular atrophy

Full-HD Methodological Framework

Given the high-dimensional origin of the omics and neuroimaging phenotypes, we developed and integrated quality control and harmonization methods of such data in the HASE software. This framework relies heavily on the partial derivatives approach, the proposed meta-analysis algorithm (developed during HD-READY) which allows for more insight into the data compared to classical meta-analysis and thus also more quality control. To illustrate this, we show that for voxel-wise analyses is it possible to generate mean gray matter density maps per cohort without access to individual-level data (Figure 2). This makes it possible to ensure that imaging processing pipelines were consistent between cohorts and all brain regions were included into analysis. During the pilot phase of Full-HD, we were able to detect, among other things, incorrect modulation of images, incorrect masking of images, incorrect normalization of phenotypes, and even in one case incorrect phenotypes themselves. Most of errors would not have been detected using the usual quality control done with classical meta-analysis. Additionally, with this framework you can reduce noise and false-positive errors, and based on such mean maps researchers can exclude phenotypes with low frequency and create a mask for the phenotype analysis space (Figure 3), the same way as it is common for genetics data and minor allele frequency. Importantly, this approach would also hold for quality control and harmonization of epigenetic data, expression data, metabolomics, and the microbiome.

Figure 2 | Gray matter density maps for three cohorts generated from the partial derivatives.

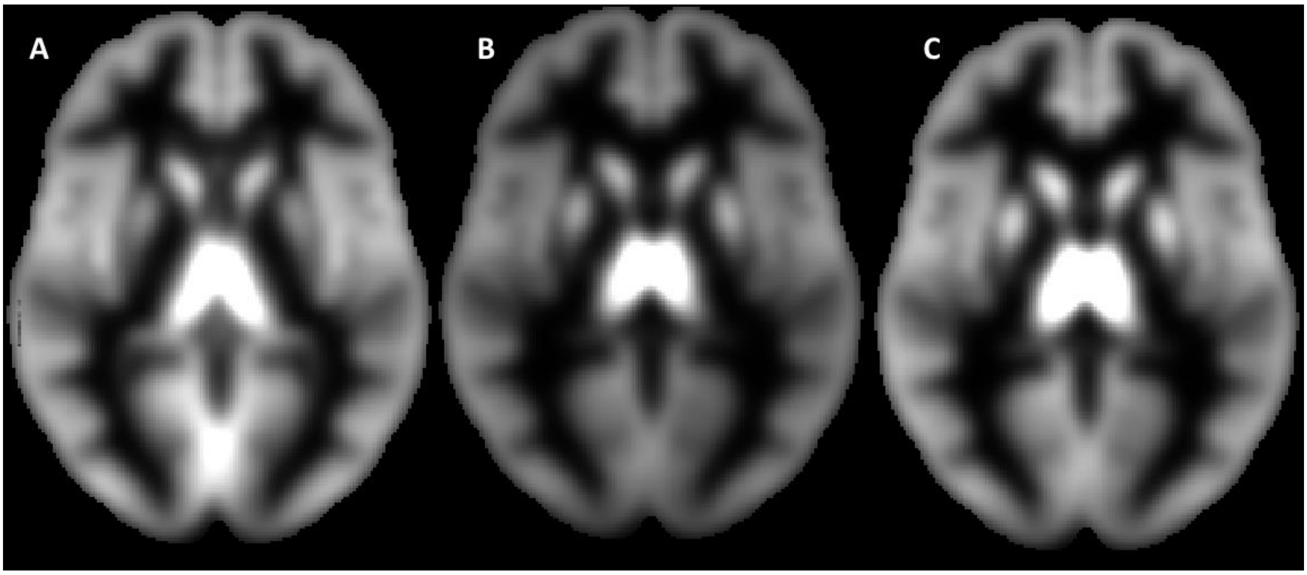
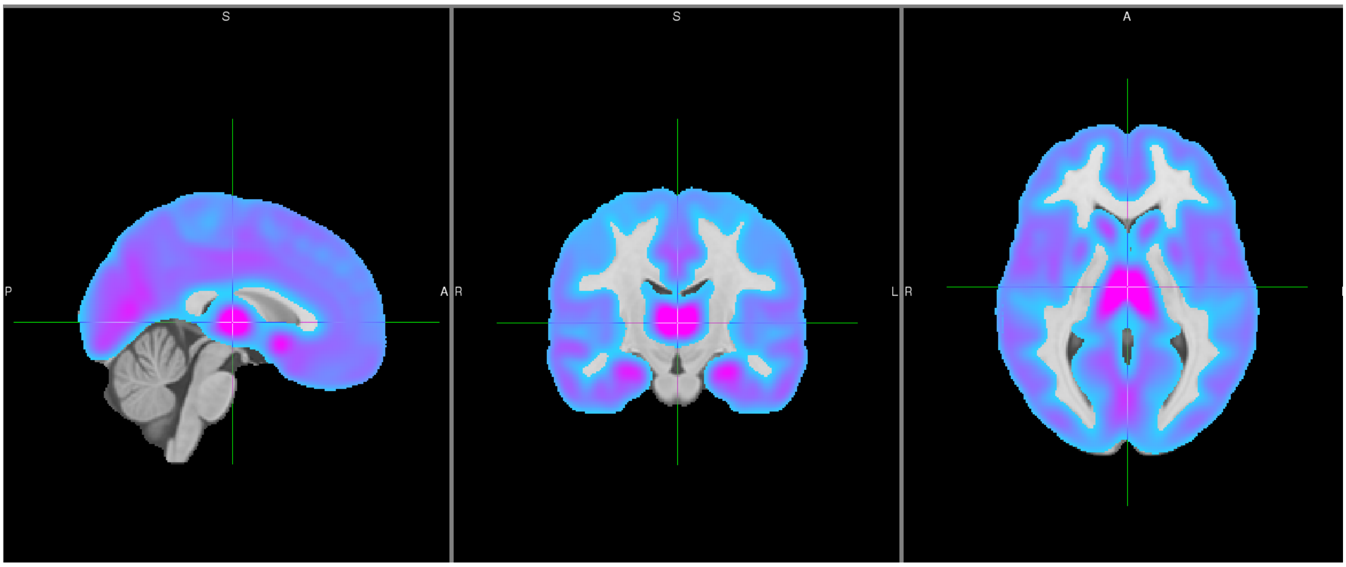


Figure 3 | Selection of harmonious phenotypes for further meta-analysis.



In the HASE software, the computational burden is already shifted almost entirely to the meta-analytical stage, making it possible for small cohort with modest computational capacity to join multi-site efforts. For the access to the resulting ultra-high-dimensional datasets, we recommend a similar solution with centralized storage of such data given the logistic burden it will place on individual sites. Two approaches are put forward. First, it is possible to store the ultra-high-dimensional data on a storage server, which, depending on the type of analysis, would require several terabytes. In this case, a database format such as hdf5, as used in HASE, will provide rapid access within such huge data. The second approach would be not to store the results of ultra-high-dimensional analyses, but rather the partial derivatives. These partial derivatives are much smaller in size, however would need some additional computation to obtain the final results. In this approach, a storage server would not be sufficient but would need to be combined with processing power for the necessary computations. An online portal providing intuitive interaction with the data would likely be most suited for everyday researchers and clinicians aiming to query the data. Those who would want to do more in-depth research with raw data can be provided with access.