HD-READY:

HIGH-DIMENSIONAL RESEARCH IN ALZHEIMER'S DISEASE

Report of a JPND Working Group on Longitudinal Cohorts

October, 2015

JPND research

EU Joint Programme – Neurodegenerative Disease Research

This document is the final report from one of ten working groups commissioned by the EU Joint Programme – Neurodegenerative Disease Research (JPND) in 2014 through a peer-reviewed call for proposals. The working groups were established to address methodological challenges preventing current population- and disease-based cohorts being further exploited for ND research. All ten reports are listed below and are available to download on the JPND website by clicking on the website link at the bottom of this page:

- **HD-READy (High-Dimensional Research in Alzheimer's Disease)**
  Coordinator: Professor M. Afran Ikram, Erasmus University Medical Centre, Rotterdam, Netherlands.

- **Harmonization and innovation of cognitive, behavioural and functional assessment in neurodegenerative dementias**
  Coordinator: Dr Alberto Costa, IRCCS Fondazione Santa Lucia, Rome, Italy.

- **NETCALS (Network of Cohort Assessment in ALS)**
  Coordinator: Professor Leonard van den Berg, University Medical Centre Utrecht, Utrecht, Netherlands

- **21st Century EURODEM**
  Coordinator: Professor Carol Brayne, University of Cambridge, Cambridge, UK

- **Multi-centre cohort-studies in Lewy-body dementia: Challenges in harmonizing different clinical and biomarker protocols**
  Coordinator: Professor Dag Aarsland, Stavanger University Hospital, Stavanger, Norway

- **Developing a methodological framework for trials in presymptomatic neurodegenerative disease – the Presymtomatic Neurodegeneration Initiative (PreNI)**
  Coordinator: Dr Jonathan Rohrer, University College London, London, UK

- **BioLoC-PD: Harmonization of biomarker assessment in longitudinal cohort studies in Parkinson's disease**
  Coordinator: Professor Daniela Berg, Hertie-Institute for Clinical Brain Research and German Center for Neurodegenerative Diseases, Tübingen, Germany

- **Dementia Outcome Measures: charting new territory**
  Coordinator: Professor Gail Mountain, University of Sheffield, Sheffield, UK

- **Body fluid biobanking of longitudinal cohorts in neurodegenerative diseases**
  Coordinator: Dr Charlotte Teunissen, VU University Medical Centre, Amsterdam, Netherlands

- **Realising the potential of cohort studies to determine the vascular contribution to neurodegeneration**
  Coordinator: Professor Joanna Wardlaw, University of Edinburgh, Edinburgh, UK

JPND Website link: http://www.neurodegenerationresearch.eu/initiatives/jpnd-alignment-actions/longitudinal-cohorts/

Table of Contents

Summary

By bringing together several large longitudinal cohort studies with imaging genetics data, this working group aims to tackle issues related to ultra-high-dimensional omics-data. Specifically, we aimed to address the research need for analytical methodology for ultra-high-dimensional data.

Achievements include:

- Partial derivatives meta-analysis: we developed a novel method to perform individual-level meta-analyses without the need for sharing individual-level data.
  - ✓ Presented at 1) the CHARGE consortium meeting in Washington DC (H. Adams), 2) the ISGC meeting in Paris (M.A. Ikram), and 3) the ENIGMA consortium in Los Angeles (H. Adams).
  - ✓ Method was well received and is being implemented by the CHARGE and ENIGMA consortia.
  - ✓ Manuscript describing the method and proof-of-principle study in 15,000 individuals from 19 cohorts of the HD-READY working group has been submitted for publication

- Efficient algorithm for ultra-high-dimensional data: a novel software package we developed for performing a large amount of computations necessary for omics data.
  - ✓ Can process 1013 computations (all brain voxels times all genetic variants) in less than two days compared to more than two years for currently available software packages
  - ✓ Suitable for linear regression analyses
  - ✓ Implementation of above-mentioned partial derivatives meta-analysis
  - ✓ Presented at the CHARGE consortium meeting in Jackson, Mississippi (June 30th 2015)
  - ✓ Manuscript in preparation for submission in September 2015

Introduction

Genetics and brain imaging are two important omics-technologies in research on neurodegenerative diseases (ND), and have undergone tremendous technological development in recent years. This development has been paralleled by an exponential growth of both imaging and genetic data. The formation of several large consortia has facilitated coordination, handling and analyses of such high-dimensional data. Some consortia, e.g. IGAP, ADGC, PERADES, focus primarily on genetics, whilst others, e.g. UNIVRSE, revolve around novel imaging markers. Only few international collaborations exist that focus on the combination of imaging and genetics, among which

CHARGE and ENIGMA are the two largest. The combination of imaging and genetics is important for ND research for several reasons. First, imaging markers can provide important endophenotypes that can aid in unravelling novel genetic variants for neurodegenerative disease. Indeed, we recently showed that MAPT is the strongest gene associated with intracranial volume – a marker of cognitive reserve relevant for AD (Ikram et al. Nat Genet 2012), whilst HRK and ASTN2 are associated with hippocampal volume (Bis et al. Nat Genet 2012). Second, relating established genes for ND with imaging markers can aid in understanding how these genes affect brain structure and function in the preclinical phase of disease. For instance, we showed that TMEM106B – the strongest common genetic factor for FTLD – affects left temporal and interhemispheric structures that are important for language (Adams et al. Biol Psychiatry 2014).

So far, the field of imaging genetics has studied either global imaging markers using hypothesis-free genetic approaches (e.g. GWAS) or candidate genes using hypothesis-free imaging approaches (e.g. voxel-based analyses). However, due to the complex nature of ND, more advanced techniques are now necessary to further investigate the brain and the genome in order to improve our understanding of the disease. Instead of looking only at the volume, it is now possible to study the shape and connectivity of the brain at a regional or voxel level. For example, persons with the same brain size, can have large differences in structural brain connection and also a different risk for AD. Similarly for genetics, single base-pair variations in the genome are now making place for the study of more complex data, such as exome, sequencing, methylation and expression data. Consequently, the field of imaging genetics will move towards combining hypothesis-free imaging analysis with hypothesis-free genetic analysis. Analytical methodology and infrastructure that address such ultra-high-dimensional data are lacking. Combining several large longitudinal cohort studies with imaging genetics data, this working group aims to tackle issues related to ultra-high-dimensional omics-data analysis.

Context

With the emergence of ultra-high-dimensional data, there is an urgent need to establish infrastructure and methods that can incorporate these new data. However, there are considerable legal, ethical, and logistic constraints in data sharing that hamper current research. Meta-analyses in consortia are therefore often performed on study-level results, but limitations of this approach are becoming increasingly apparent with more complex analyses. Although pooling all participant data in a single location has been successfully done by others (e.g. the Psychiatric Genetics Consortium), this is unrealistic and even prohibited for most cohort studies, including those participating in our working group. Given these constraints, our working group aims to address the analytical methodology for ultra-high-dimensional data.

Studies that are willing and able to share individual-level data are scarce, and in combination would not reach the needed sample size even remotely. Study-level meta-analysis is generally agreed to by cohort studies, but its results are less accurate, not flexible, and obscure possible errors in the analysis that was performed. Given these considerable disadvantages of the two classical data-sharing strategies, there is an urgent need for a meta-analytical approach, which 1) provides effect estimates that are equivalent to individual-level meta-analysis; 2) can be shared without disclosing identifiable participant information; 3) is easily applied to large imaging and genetic datasets. This Working Group aims to develop analytical approaches that address these issues.

Terminology

Biomarkers objectively and accurately measured indicators of a person's medical state
GWAS genome-wide association study
MRI magnetic resonance imaging

Methods

First HD-READY meeting on 26-27 October 2014 in Rotterdam

- As per the JPND committee's recommendations, the meeting focused primarily on Aim 1 of the proposal: Developing a framework and methodology for performing high-dimensional analysis (combining imaging and genetics)
- The workshop was set-up into two halves. In the morning plenary presentations were given. In the afternoon the group split in 2 to discuss separately the framework and methodology.
- Hieab Adams presented the 'partial derivative meta-analysis', a novel method to perform individual-level meta-analyses without the need for sharing individual-level data. Within the smaller group in the afternoon, a plan of action was agreed to expand this method. Specifically, 1) the proof-of-concept analysis of FTO/BMI/Brain volume will be finalized and written into a paper, 2) scripts will be written to automate this methodology in order to facilitate high-throughput, 3) further generalization of the technique to include logistic regression.
- Gennady Roshchupkin has been working on a novel technique to visualize hypothesis-free imaging data using a 'Manhattan' brain plot. The proof-of-concept of this tool is close to being submitted and subsequently will be made freely available. Other HD-READY cohorts can then implement the tool for their own datasets.
- In the second subgroup, M. Arfan Ikram led the discussion on hypothesis-free imaging analyses (i.e. voxel-based analyses). The idea was to brainstorm on how imaging studies should be performed if standards used in genetics would be applied. This pertains to 1) sample size; 2) statistical thresholds and 3) replication. The idea is the write this up as a 'recommendation' to be submitted to JPND. If there is sufficient subject matter and people are interested this could also serve as basis for a position manuscript or commentary on 'what can neuroimaging learn from neurogenetics?'.

Second HD-READY meeting on 22 February 2015 in Rotterdam

- The second HD-READY meeting followed a similar format as the first meeting. Importantly, we also invited the three members of the External Advisory Committee to provide input and assess our progress. Specific topics discussed at the meeting were:
    - To show proof of principles for the various novel methodologies introduced in the first meeting. Hieab Adams and Gennady Roshchupkin both presented results for the novel meta-analysis tools and the rapid analysis of high-dimensional data.
    - We also invited several investigators from the ENIGMA consortium to provide expert talks on similar work that they have been conducting. Derek Hibar, Neda Jahanshad, and Paul Thompson presented their work on shared genetics, diffusion tensor imaging genetics, and an overview of the field, respectively.
    - We had a session in which members of other JPND working groups presented their aims and progress. Joana Wardlaw showed some work done in the vascular brain disease working group and Arfan Ikram updated the group about the work from the working group led by Carol Brayne,
    - The final session was dedicated to formulate main conclusions of this working group that formed the basis for the present report. Also, we discussed next steps for this consortium as a whole. Given the size of the consortium, we would not be eligible to submit a proposal in the current JPND call. Nevertheless, the work and methodology derived in this working group will be used in the BRIDGET-proposal that has been submitted to JPND.

The JPND HD-READY working group at the second meeting in Rotterdam

AGREED GUIDELINES

With the emergence of ultra-high-dimensional data, there is an urgent need to establish infrastructure and methods that can incorporate these new data. However, there are considerable legal, ethical, and logistic constraints in data sharing that hamper current research. Meta-analyses in consortia are therefore often performed on study-level results, but limitations of this approach are becoming increasingly apparent with more complex analyses. Although pooling all participant data in a single location has been successfully done by others (e.g. the Psychiatric Genetics Consortium), this is unrealistic and even prohibited for most cohort studies, including those participating in our working group. Given these constraints, the HD-READY working group proposes the '4S Principles' that can serve as guidelines for analyzing ultra-high-dimensional data.

HD-READY - 4S Principles

SETTING:
For correct application of any methodology and correct interpretation of any results, it is important to realize the setting from which the sample and data originate. This has important impact on generalizability of findings. Given that in genetics studies of imaging markers, confounding by disease status is unwanted, it is recommended to use population-based samples for such work. The advantages of such samples are: no pre-selection of disease status, homogeneity of samples, generalizability to the source population, and usually longitudinal measures. At the same time, given that we are seeking subtle genetic effect on equally subtle imaging markers, a major bottleneck for any study is power. In this regard, increase of sample size is pivotal. In the balance between increasing sample size by adding non-population-based studies (thus increasing heterogeneity), at this moment the increase of sample size is preferred.
Empirical evidence of the fact that larger power is preferable than homogeneous samples comes from collaborative work between CHARGE (mostly population-based samples) and ENIGMA (mostly non-population-

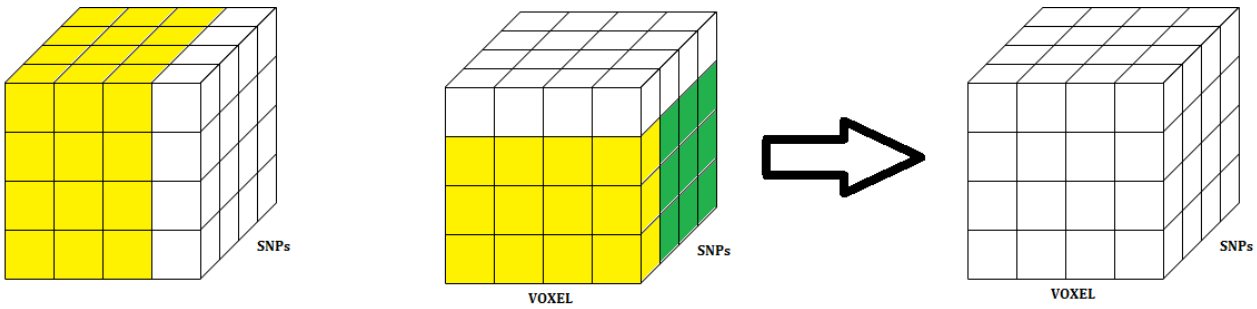
based samples), which in joint meta-analysis have identified various novel variants for MRI-markers (i.e. intracranial volume and hippocampus volume).

SPEED:

A second important consideration in ultra-high-dimensional imaging genetics is the speed of processing and calculations. In its most extreme form a voxel-based GWAS would yield an analytical framework encompassing 1.5 million x 9 million regressions. Using conventional analytical software and computational infrastructure such effort would take >400 years per sample. Several optimization algorithms have been described in the literature, but the gain has merely resulted in a time-requirement of 2 years.

Under the umbrella of HD-READY we are currently working on a novel method that dramatically reduced processing time by removing redundant computations. Our initial work suggested that of all the 1.5 million x 9 million computations, >95% are redundant and therefore only need to be done once. Additionally, we have systematically investigated libraries used for these computations and file formats for the storage of the genetic and imaging data. We found that optimizations from other fields have great potential for the analysis of high-dimensional data in research. We have now completed a proof-of-concept study in which the full analysis of all voxels and genetic variants was performed in 2 days. Furthermore, partial derivatives meta-analysis is implemented as an additional feature. Our current recommendation therefore is that, for high-dimensional data, this novel method is preferable with respect to computational efficiency and analytical superiority.

3D representation of necessary computations for a voxel-wise genome-wide association study, and how the identical results can restricted using only a subset of the computations

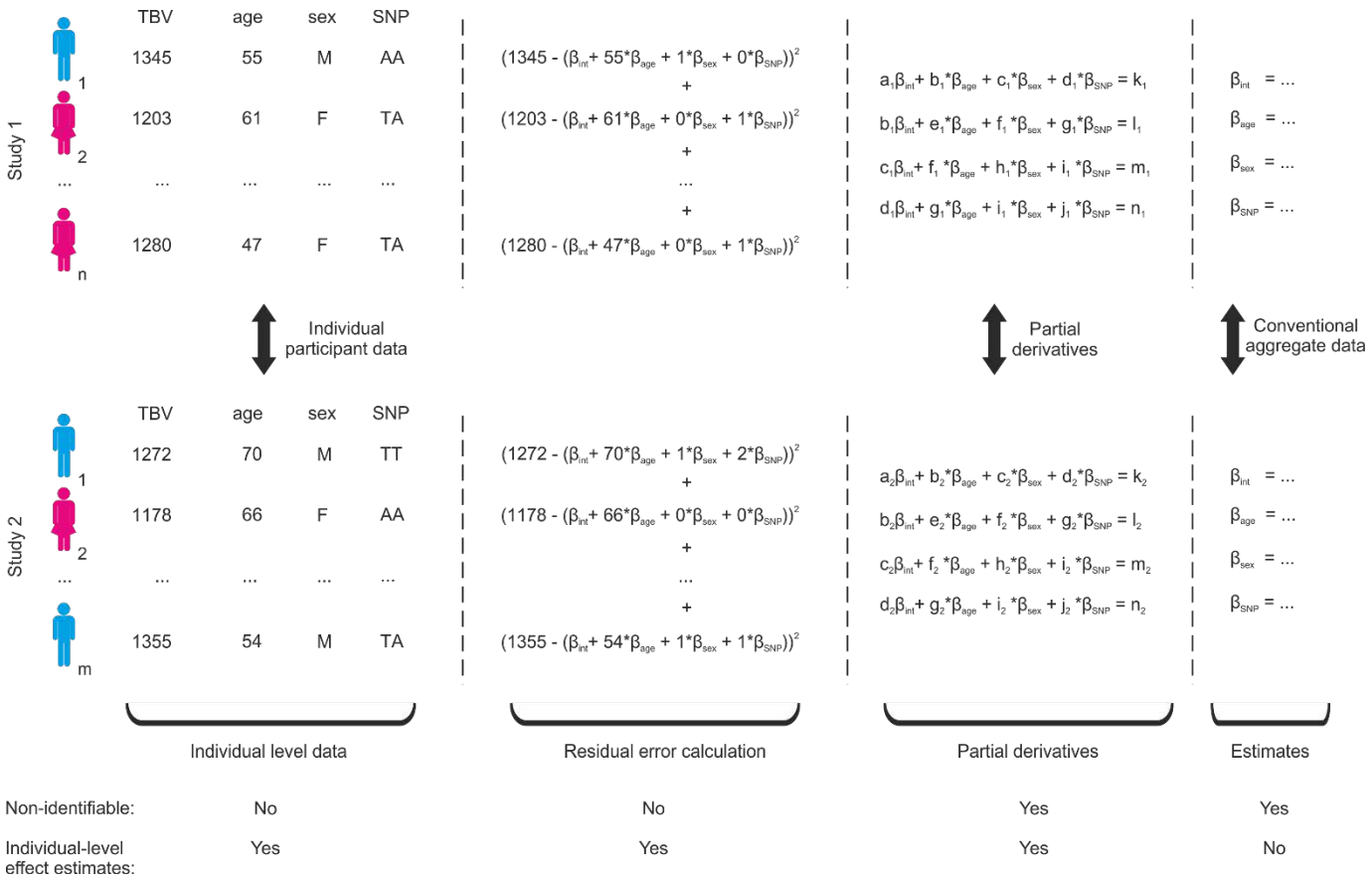


SHARING:

In contrast to above-mentioned constraints in imaging genetics, which are both scientific in nature, perhaps the most important consideration is actually non-scientific, but rather logistic: several studies have restrictions in the type and amount of data that can be shared with external collaborators. Especially in genomics, this has resulted in some studies being left out of consortia, or only participating in a subset of all ongoing work. An alternative has been to use approximate methods that do not require sharing of individual data, but instead approximate individual level analysis using study-level meta-data. An ideal scenario would combine the advantages of both: perform individual-level meta-analysis without the need of individual-level data. Therefore, we developed partial derivatives meta-analysis, a novel meta-analytical method that we developed specifically to overcome

the limitations of classical meta-analysis. Our method allows for true multi-study joint analysis without the need for exchanging the actual participant data. Instead of sharing the individual-level data (e.g. age, sex, disease status, genotype) or the study-level effect estimates (e.g. beta coefficients, odds ratios), our method uses intermediate statistics: the partial derivatives. These partial derivatives cannot be traced back to individual-level data, but do provide effect estimates that are mathematically equivalent to a pooled analysis. We applied this method to data from almost 15,000 participants as a proof-of-principle study where we investigated the relation between obesity and brain volume to illustrate its benefits. The method is described in detail elsewhere [Adams et al. Submitted].

Comparison between classical methods of meta-analysis and our novel method, partial derivatives meta-analysis



STATISTICS:

Even after following the above 3 recommendations, a major issue remains in the correct interpretation of the results. A typical voxel-based GWAS (1.5 million x 9 million) will yield 1.35 x 10^{13} statistical tests and p-values. Such an enormous amount of output comes with its own set of challenges. These include storage, transport, and visualization. Within HD-READY we focused on an additional challenge: what threshold to use for statistical

significance? Answering this question is not only important for correct interpretation of the obtained results, but also functions as a basis for proper replication efforts.

Currently, several tentative thresholds have been suggested in the literature. Thompson et al reported 5×10^{-12} as threshold and based their MRI simulation using the standard and widely used threshold of 5×10^{-8} for a single trait GWAS and correcting for around 10,000 tests in a three-dimensional image. As they rightly pointed out their threshold was an approximation, because of the use of various assumptions and approximation, and further refinement of this threshold is needed. We calculated that there were about 100,000 independent tests for brain imaging data from the Rotterdam Study and ERF study, and thus suggest to use 5×10^{-13} as the significance threshold for brain-wide and genome-wide association studies.

HD-READY 4S Principles	
SETTING	Population-based studies have more homogeneity than clinical samples, but larger sample sizes (by combining both types of studies) are more important
SPEED	Use novel method for high-dimensional data
SHARING	Use partial derivatives for sharing aggregate data that produces gold standard results and protects participant privacy
STATISTICS	Multiple testing thresholds for genome-and-brain-wide associations scans have yet to be validated; we propose $p < 10^{-14}$

Contributors

Name	Institute	Country
Arfan Ikram	Erasmus MC	the Netherlands
Alexander Teumer	Greifswald University	Germany
Bernard Mazoyer	Univ of Bordeaux	France
Christiane Reitz	Columbia University	USA
Claudia Satizabal	Boston University	USA
Cornelia M. van Duijn	Erasmus MC	the Netherlands
Derrek Hibar	USC	USA
Ganesh Chauhan	Paris 7 University	France
Gennady Roshchupkin	Erasmus MC	the Netherlands
Hans Grabe	Greifswald University	Germany
Helena Schmidt	University of Graz	Austria
Hieab Adams	Erasmus MC	the Netherlands
Jeroen van der Grond	Leiden Univ MC	the Netherlands
Joanna Wardlaw	Univ of Edinburgh	UK

Kamran Ikram	National University Singapore	Singapore
Meike Vernooij	Erasmus MC	the Netherlands
Myriam Fornage	University of Houston	USA
Natalia Rost	Harvard Medical School	USA
Neda Jahanshad	USC	USA
Paul Nyquist	Johns Hopkins University	USA
Paul Thompson	USC	USA
Reinhold Schmidt	University of Graz	Austria
Saima Hilal	National University Singapore	Singapore
Sarah Madsen	USC	USA
Stéphanie Debette	Paris 7 University	France
Sven van der Lee	Erasmus MC	the Netherlands
Sudha Seshadri	Boston University	USA
Ton de Craen	Leiden Univ MC	the Netherlands
Vincent Chouraki	Boston University	USA
Wiro Niessen	Erasmus MC	the Netherlands

Acknowledgments

Arfan Ikram was supported by ZonMW grant number 916.13.054. Hieab Adams was supported by the Van Leersum Grant of the Royal Netherlands Academy of Arts and Sciences.

WWW.JPND.EU

JPND research

EU Joint Programme – Neurodegenerative Disease Research