Predictive Analytics of Big Neuroscience Data

Ivo D. Dinov, Nina Zhou, Syed Husain, Alexandr Kalinin, Yi Zhao, and Simeone Marino

Statistics Online Computational Resource (SOCR), Departments of Health Behaviour and Biological Sciences (HBBS) and Computational Medicine and Bioinformatics (DCMB), Michigan Institute for Data Science (MIDAS), University of Michigan, Ann Arbor, MI 8109, USA

Abstract:
This work present some of the Big neuroscience data research and education challenges and opportunities. Specifically, I identify the core characteristics of complex neuroscience data, discuss strategies for data harmonization and aggregation, and show case-studies using large data of normal and pathological cohorts. Examples of the demonstrated techniques include *DataSifter*, which enables secure sharing of sensitive data, compressive big data analytics, which facilitates inference on multi-source heterogeneous datasets, and model-free prediction providing forecasting of clinical features or derived computed phenotypes. Simulated data as well as clinical data (e.g., UK Biobank (UKBB), Alzheimer's Disease Neuroimaging Initiative (ADNI), and amyotrophic lateral sclerosis (ALS) case-studies) are used for testing and validation of the techniques. In support of *open-science*, result reproducibility, and methodological improvements, all datasets, statistical methods, computational algorithms, and software tools are freely available online.

Keywords:
Big Data; Model-based analytics; Model-free inference; Neurodegenerative disorders; Data science; Open-science

1. Introduction:

This paper aims to present some of the contemporary Big neuroscience data challenges, provide examples of solutions for specific problems, and identify research, computational, and educational opportunities. We will begin by defining data science and predictive analytics and examining the common characteristics of Big datasets. Focusing on several driving biomedical and health challenges, we will pinpoint some concrete barriers to data sharing. We will briefly review two complementary strategies to enable data computing on sensitive information, ε-differential privacy (Dwork 2009) and homomorphic encryption (Gentry 2009). Then, we will describe a recently introduced technique for statistical obfuscation of sensitive data (*DataSifter*) and demonstrate its approach to balancing data security and data-utility (Marino, Zhou et al. 2018). We will conclude by examining three biomedical and health applications using neurodegenerative aging disorders, paediatric pathological brain development, and exploratory census-like population neuroscience.

Figure 1 shows a schematic that illustrates the relation between census-like population-based view of natural processes (left), their Big Data proxy representation (middle), and classical (small) sampling based process description. By examining many dozens of complex biomedical and health case-studies we identified the common characteristics of Big Data (Dinov 2018), **Table 1**.

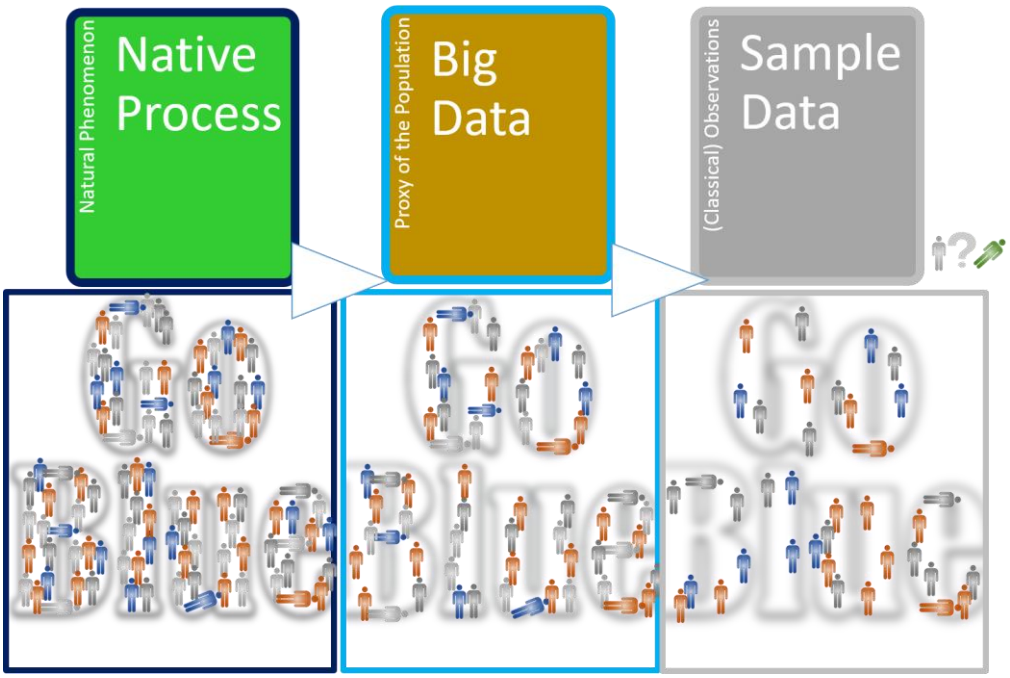


Figure 1: Schematic of the relation between native processes (left), their Big Data representations (middle), and traditional sampling based process characterization. Note that (1) the ideal population view of the process is often unobservable and intractable, the Big Data proxy of the process often requires substantial data management, harmonization, aggregation, preprocessing and wrangling before it can be analysed, and (3) the sample data may facilitate rapid and effective data analytics, but may also represent a limited view of the entire process.

Table 1: Common characteristics of Big biomedical and healthcare datasets.

Dimensions of Big Data	**Properties and Tool specifications**
Size	Harvesting and management of vast amounts of data
Complexity	Wranglers for dealing with heterogeneous data
Incongruency	Tools for data harmonization and aggregation
Multi-source	Transfer and joint modeling of disparate elements
Multi-scale	Macro to meso to micro scale observations
Time	Techniques accounting for longitudinal patterns in the data
Incomplete	Reliable management of missing data

2. Methodology:

There are a few complementary strategies that enable scientific computing on sensitive datasets. Examples of these include ε-differential privacy (Dwork 2009), homomorphic encryption (Gentry 2009), and statistical obfuscation via *DataSifter* (Marino, Zhou et al. 2018). Below we review each of these techniques.

2.1 ε-differential privacy (ε-DP)

ε-DP provides a mechanism to mine information in databases without compromising privacy. By estimating the ttheoretical limits on the balance between information utility and risk of sharing data, this technique enables data governors to quantify the potential risks of information re-identification. However, it is difficult to apply for unstructured, skewed, or categorical data (Dwork 2009).

Assume we have a dataset including measurements of the following features: $\{C_1, C_2, \ldots, C_k\}$, which can be categorical or numerical. Relational databases (DBs) store lists of cases $\{x_1, x_2, \ldots, x_n\}$, $x_i \in C_1 \times C_2 \times \ldots \times C_k$, $1 \le i \le n$. ε-Differential privacy relies on adding noise to the data in the database, which adds protection against reidentification of individual records. An algorithm f is called *ε-differentially private* if for all possible inputs (datasets or DBs) D_1, D_2 that differ on a single record and all possible f outputs, y, the probability of correctly guessing D_1 knowing y is not significantly different from the corresponding probability of D_2 given y. In other words,

$$\frac{P(f(D_1) = y)}{P(f(D_2) = y)} \le e^{\varepsilon}, \quad \forall y \in Range(f).$$

Clearly the small positive number, $\varepsilon > 0$ and $e^{\varepsilon} \sim 1$, controls the level of uncertainty about reidentification of the source data (D_1 or D_2) from the known observation, y. The *global sensitivity* of f is the smallest number $S(f)$, such that $\forall D_1, D_2$ that differ on at most one element $\|f(D_1) - f(D_2)\|_1 \le S(f)$. There are many differentially private algorithms, e.g., random forests, decision trees, k-means clustering, etc. For instance, if $f: D = DB \to R^m$, the algorithm outputting $f(D) + (\eta_1, \eta_2, \ldots, \eta_m)$, with $\eta_i \in Laplace\left(\mu = 0, \sigma = \sqrt{2}\frac{S(f)}{\varepsilon}\right)$, $\forall i$ is ε-differentially private.

2.2 Fully-Homomorphic Encryption (FHE)

FHE security is based on preprocessing the data by encryption to allow subsequent program execution and data-driven inference using the encrypted information (Gentry 2009). As a result, the process outputs are encrypted and their interpretation requires ability to *decrypt* the information following the data analytics. It represents an elegant and powerful mathematical framework for bijective (encoding/decoding) processing and analytics. Albeit, it is very fast, FHE has some limitations, e.g., deriving the $\boldsymbol{f'}$ – commutative analytic evaluators – is never a trivial task and requires close cooperation between data governor and data user. **Figure 2** shows schematically the process of data analytics using fully-homomorphic encryption.

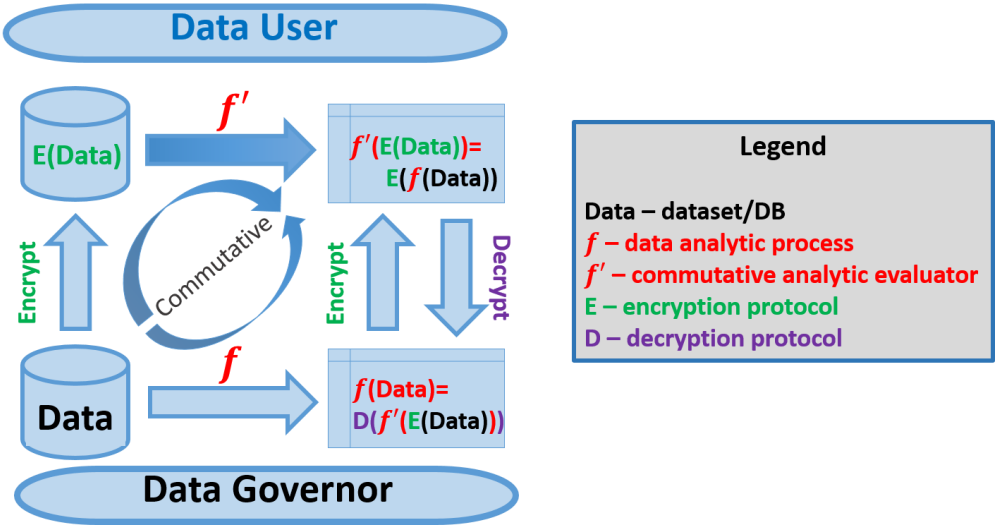


Figure 2: Data analytics via fully-homomorphic encryption.

2.3 *DataSifter* Statistical Obfuscation

The process of data-masking using statistical obfuscation is the core of the *DataSifter* technique. It combines artificial random missingness with partial information alterations using data swapping within subjects' neighbourhoods. These operations have minimal impact on the joint distribution of the obfuscated (sifted) output data as the controlled rate of missingness is introduced completely at random and nearest neighbourhoods tend to have consistent distributions. The *DataSifter* algorithm preserves the bulk of the total data energy of the original data in terms of conserving the overall distribution of the original data features. Simultaneously, the method obfuscates the individual cases sufficiently to protect against the risks of subject re-identification. The *DataSifter* technique includes

several user-controlled parameters that allow the data governor the flexibility to control the level of obfuscation, trading privacy protection and preservation of signal energy (Marino, Zhou et al. 2018). **Figure 3** shows a schematic of the *DataSifting* protocol.

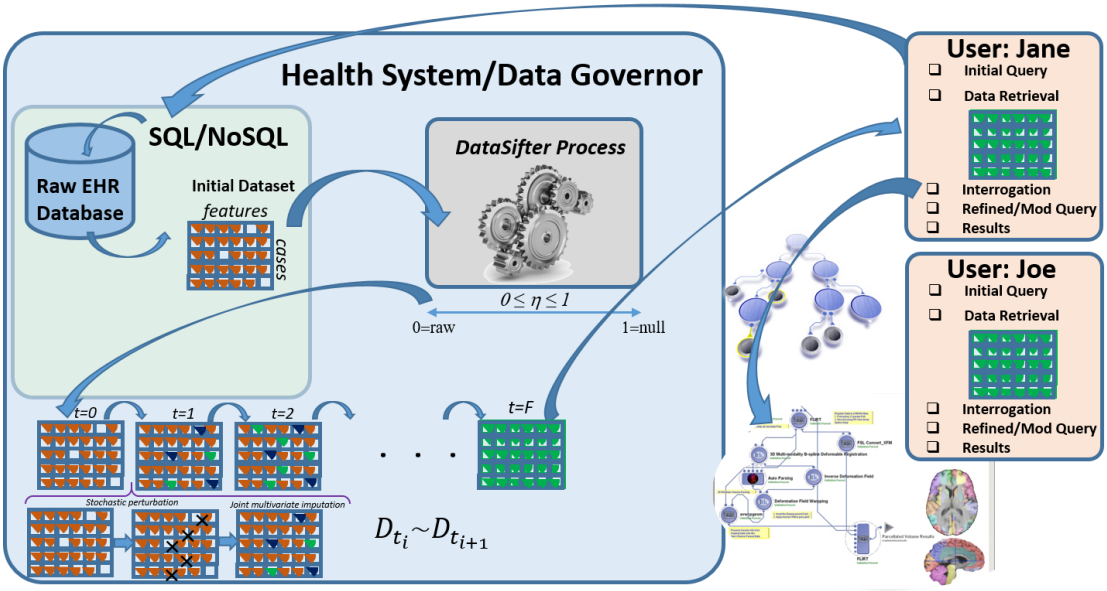


Figure 3: Summary of the *DataSifter* protocol.

Figure 4 illustrate the validation results of applying the DataSifter to a specific clinical case-study. In this case we obfuscated a large Autism Brain Imaging Data Exchange (ABIDE) dataset including 1,098 volunteers and 2,400 features (http://fcon_1000.projects.nitrc.org/indi/abide) (Di Martino, Yan et al. 2014, Torgerson, Quinn et al. 2015). The results include the Percent of Identical Feature Values (PIFV), vertical axis, for different *DataSifter* obfuscation levels. Each box represents all subjects in the ABIDE sub-cohort and random forest prediction of a specific binary clinical outcome - autism spectrum disorder – (ASD) status (ASD vs. control).

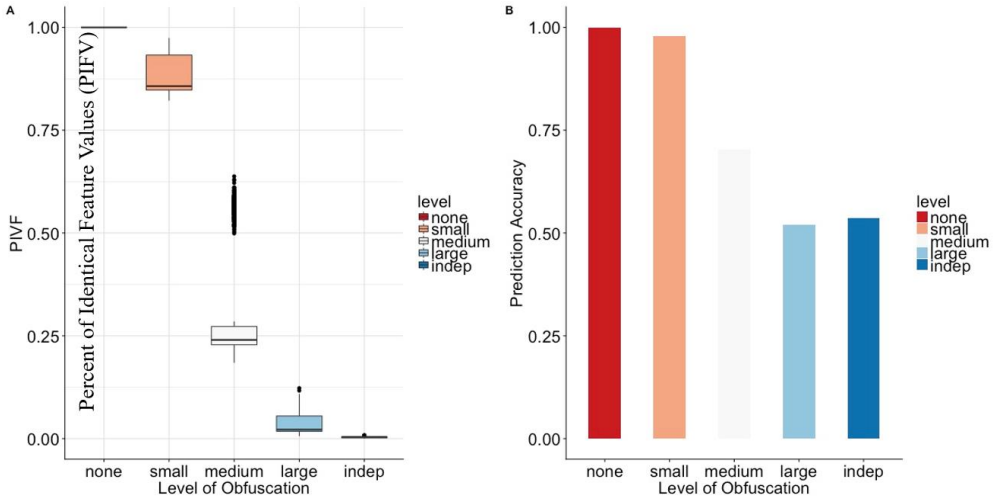


Figure 4: DataSifter obfuscation – trade-offs between privacy protection and preservation of data utility.

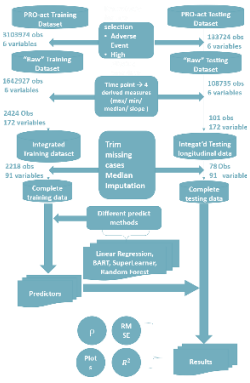
In addition, we use established model-based and model-free techniques to interrogate the data (Dinov 2016, Dinov 2016, Dinov 2018, Gao, Sun et al. 2018, Kalinin, Allyn-Feuer et al. 2018, Marino, Xu et al. 2018, Tang, Gao et al. 2018, Zhao, Matloff et al. 2018). These include both confirmatory (hypothesis driven) and exploratory (visual analytics) inferential techniques to extract knowledge, identify patterns, forecast trends, and forecast univariate outcomes of interest and derived computed phenotypes.

3. Results:

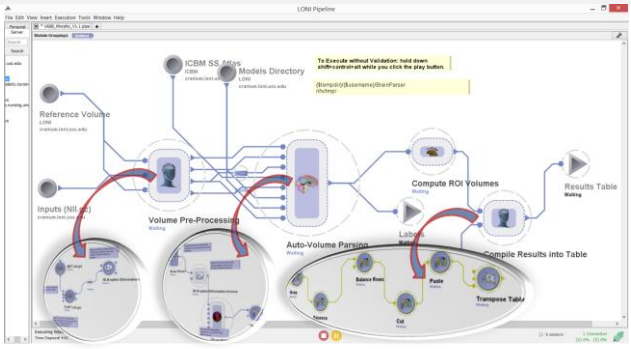
Open-science relies heavily on data sharing, findability, accessibility, interoperability, and reusability (FAIR) (Wilkinson, Dumontier et al. 2016), open-source development (Feller and Fitzgerald 2002), and transdisciplinary cooperation (Kreps and Maibach 2008, Dinov 2018). **Figure 5** presents some examples of recent results illustrating the power of advanced mathematical modelling techniques, statistical inferential methods, and machine learning strategies to analyse complex, multisource, heterogeneous, and incomplete datasets.

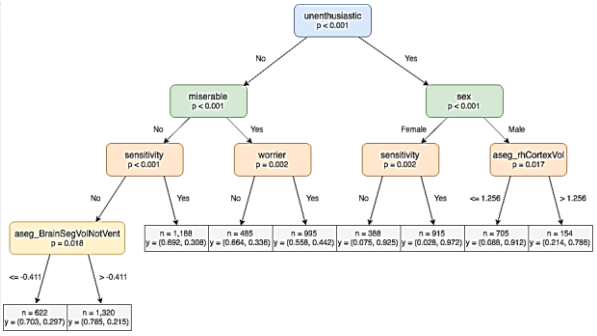
The SOCR Data Dashboard enables aggregation of multisource data and visual data query and analytics (Husain 2015).

Cluster	Consistency	Variance	Cluster-Size	Silhouette
1	1	0	565	0.58
2	0.99	0.02	427	0.63
3	0.96	0.05	699	0.5
4	0.99	0.02	733	0.5



Amyotrophic Lateral Sclerosis (ALS) study aiming to predict disease progression using clinical and lab test information of 2,424 participants and over 2,400 features (Huang, Zhang et al. 2017, Tang, Gao et al. 2018).





A study examining over 10,000 participants in the UKBB cohort identified deep phenotypic traits in the population related to mental health using unsupervised machine learning methods (Zhao, Zhao et al. 2019). The left panel above shows the automated end-to-end computational pipeline workflow deriving thousands of brain morphometric features. The panel on the right shows a decision tree illustrating a simple clinical decision support system providing machine guidance for identifying depression feelings based on categorical variables and neuroimaging biomarkers. Each terminal node, includes the percentage of subjects being labelled as "no" and "yes", in this case, answering the question "Ever depressed for a whole week." The p-values listed at branching nodes indicate the significance of the corresponding splitting criterion.

Figure 5: Examples of recent Big health data analytic studies.

4. Discussion and Conclusion:

There are many remaining data science "open problems" including establishing the fundamentals of data representation, modelling, and analytics, quality control and data value metrics, and effectively strategies for data wrangling, harmonization, aggregation, and joint understanding. There also are terrific opportunities for scientific discoveries, basic science developments, ubiquitous range of applications, development of effective educational resources, and designing learning modules to engage a wider cross-section your investigators. All these activities demand substantial community, institutional, state, federal, international, and philanthropic support to advance data analytic methods, enhance the computing infrastructure, train and support students and fellows, and tackle the $Kryder\ Law \gg Moore\ Law$ trend (Dinov 2014).

Acknowledgements: This research is supported in part by NSF grants 1734853, 1636840, 1416953, 0716055 and 1023115; NIH grants P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, UL1TR002240; and the Elsie Andresen Fiske Research Fund.

References

1 Di Martino, A., C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer and M. Dapretto (2014). "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism." Molecular psychiatry **19**(6): 659.

2 Dinov, I. (2016). "Methodological Challenges and Analytic Opportunities for Modeling and Interpreting Big Healthcare Data." GigaScience **5**(12): 1-15.

3 Dinov, I. (2018). Data Science and Predictive Analytics: Biomedical and Health Applications using R, Springer International Publishing.

4 Dinov, I., Petrosyan, P, Liu, Z, Eggert, P, Hobel, S, Vespa, P, Woo, Moon S, Van Horn, JD, Franco, J, and Toga, AW. (2014). "High-Throughput Neuroimaging-Genetics Computational Infrastructure." Frontiers in Neuroinformatics **8**(41): 1-11.

5 Dinov, I. D. (2016). "Volume and value of big healthcare data." Journal of Medical Statistics and Informatics **4**(1): 1-7.

6 Dwork, C. (2009). The differential privacy frontier. Theory of Cryptography Conference, Springer.

7 Feller, J. and B. Fitzgerald (2002). Understanding open source software development, Addison-Wesley London.

8 Gao, C., H. Sun, T. Wang, M. Tang, N. I. Bohnen, M. L. T. M. Müller, T. Herman, N. Giladi, A. Kalinin, C. Spino, W. Dauer, J. M. Hausdorff and I. D. Dinov (2018). "Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease." Scientific Reports **8**(1): 7129.

9 Gentry, C. (2009). A fully homomorphic encryption scheme, Stanford University.

10 Huang, Z., H. Zhang, J. Boss, S. A. Goutman, B. Mukherjee, I. D. Dinov, Y. Guan and A. L. S. C. T. C. for the Pooled Resource Open-Access (2017). "Complete hazard ranking to analyze right-censored data: An ALS survival study." PLOS Computational Biology **13**(12): e1005887.

11 Husain, S., Kalinin, A, Truong, A, Dinov, ID (2015). "SOCR Data dashboard: an integrated big data archive mashing medicare, labor, census and econometric information." Journal of Big Data **2**(13): 1-18.

12 Kalinin, A. A., A. Allyn-Feuer, A. Ade, G.-V. Fon, W. Meixner, D. Dilworth, S. S. Husain, J. R. de Wett, G. A. Higgins, G. Zheng, A. Creekmore, J. W. Wiley, J. E. Verdone, R. W. Veltri, K. J. Pienta, D. S. Coffey, B. D. Athey and I. D. Dinov (2018). "3D Shape Modeling for Cell Nuclear Morphological Analysis and Classification." Scientific Reports **8**(1): 13658.

13 Kreps, G. L. and E. W. Maibach (2008). "Transdisciplinary science: The nexus between communication and public health." Journal of Communication **58**(4): 732-748.

14 Marino, S., J. Xu, Y. Zhao, N. Zhou, Y. Zhou and I. Dinov (2018). "Controlled Feature Selection and Compressive Big Data Analytics: Applications to Biomedical and Health Studies." PLoS Bioinformatics **13**(8): e0202674.

15 Marino, S., N. Zhou, Y. Zhao, L. Wang, Q. Wu and I. D. Dinov (2018). "HDDA: DataSifter: statistical obfuscation of electronic health records and other sensitive datasets." Journal of Statistical Computation and Simulation **HDDA**: 1-23.

16 Tang, M., C. Gao, S. A. Goutman, A. Kalinin, B. Mukherjee, Y. Guan and I. D. Dinov (2018). "Model-Based and Model-Free Techniques for Amyotrophic Lateral Sclerosis Diagnostic Prediction and Patient Clustering." Neuroinformatics.

17 Torgerson, C., C. Quinn, I. Dinov, Z. Liu, P. Petrosyan, K. Pelphrey, C. Haselgrove, D. Kennedy, A. Toga and J. Van Horn (2015). "Interacting with the National Database for Autism Research (NDAR) via the LONI Pipeline workflow environment." Brain Imaging and Behavior: 1-15.

18 Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos and P. E. Bourne (2016). "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data **3**.

19 Zhao, L., W. Matloff, K. Ning, H. Kim, I. D. Dinov and A. W. Toga (2018). "Age-Related Differences in Brain Morphology and the Modifiers in Middle-Aged and Older Adults." Cerebral Cortex: bhy300-bhy300.

20 Zhao, Y., L. Zhao, N. Zhou, Y. Zhao, S. Marino, T. Wang, H. Sun, A. Toga and I. Dinov (2019). "Predictive Big Data Analytics using the UK Biobank Data." Scientific Reports **in press**.