Distribution Agreement

____ _______________

Signature of Student Date

Big Data Analytics in Bioinformatics: Microservices Architecture in the Cloud - Alzheimer's Surveillance Re-envisioned

BY

Brooke E. Rivera, MBA

Degree to be awarded: M.P.H.

Executive MPH

____ ________________

Signature of Student Date

____ ________________

Chair, Executive MPH Program Date

____ ________________

Program Chair, Executive MPH Program Date

ABSTRACT

Research indicates that over eighty percent of brain disorders are associated with genomics defects in conjunction with environmental factors and epigenetic phenomena[i]. There is a significant body of evidence that suggests that the confluence of next-generation sequencing (NGS), cloud computing, and advanced analytics are producing a paradigm shift in our understanding of disease etiology, development and treatment, challenging the traditional surveillance model.[ii] [iii] [iv] [v] [vi] Advances in computing power and "big data" provide the engine for mathematical and statistical techniques by which disparate datasets can be synthesized and analyzed.[vii] Whole-genome sequencing (WGS) offers precision resolution in orders of magnitude over earlier genotyping methods, transforming our approach to monitor, control, and prevent diseases in both epidemic and endemic contexts.[viii] In addition, the increased availability of dense data characterizing the substrate population and the development of advanced computation and analytical tools to organize and interpret these large datasets broadens the potential for application of such data to high-resolution epidemiological problems.[ix] In order to harness this potential, global surveillance systems and initiatives must be strengthened, tightly coordinated, and judiciously harmonized to inform global strategies, monitor the effectiveness of public health interventions, and detect new threats. *Esperanza 3.0***,** an integrated cloud-based multi-disciplinary surveillance platform, will offer the foundation for this united global response.

Esperanza 3.0 is envisioned to be a highly secure containerized microservices platform which will host an international repository for Alzheimer's case notifications, electronic lab reports, and imaging records for participating nation-states. Cloud computing offers unprecedented access to highly performant, supercomputing resources historically reserved for military and classified research facility operations – allowing us to leverage the scalability, elasticity, enriched analytics capacity, and computing power. Deep drilling and cluster analysis

will be introduced for longitudinal research studies, monitoring, control and intervention. This platform will be architected to modernize, optimize, and ultimately revolutionize chronic disease surveillance by integrating modern data visualization and advanced analytics.

BACKGROUND

In addition to the human suffering caused by the disease, Alzheimer's is creating an enormous strain on the health care system, families and caregivers and the federal budget. Alzheimer's is a progressive brain disorder that damages and eventually destroys brain cells, leading to a loss of memory and neuro-cognitive function. Ultimately, Alzheimer's is fatal. Currently, Alzheimer's is the sixth leading cause of death in the United States and the only one of the top ten without a means to prevent, cure or slow its progression. Over five million Americans are living with Alzheimer's, with 200,000 under the age of 65. It is anticipated that over the next 40 years, the escalating national epidemic (based on the current trajectory), Alzheimer's, and other forms of dementia, will cost the American society alone, more than $20 trillion.[x] With the advent of a secure, cloud-based surveillance platform, data scientists may collaborate to synthesize disparate data sets and longitudinal health records (e.g., chronic disease indicators, demographics, genomics, epidemiological, environmental, laboratory, geo-spatial, pharmaceutical) to:

- Integrate clinical, epidemiological, and genomics data to support translational research, transforming genetics-related discoveries into practical health care application.
- Provide empirical evidence to expand successful health care programs and interventions for caregivers and promote their adoption into widespread practice.
- Develop innovative tools, visualizations, and reports to improve access to data and share findings about cognitive decline and its consequences, the prevalence and characteristics of caregivers and care recipients, and the physical and mental health of older Americans.

- Support efforts to increase awareness about Alzheimer's and other dementias.
- Track progress toward Healthy People 2020 objectives for dementia and older adults.[xi]
- Work with partners to develop and share strategies that help older adults with a cognitive impairment to remain active, independent, and involved in their community.

At present, possibilities for prevention, early detection, and treatment of these diseases are limited. Predictive and diagnostic genetic testing is available only for limited forms of Alzheimer's, and even screening, pre-symptomatic testing, or clinical diagnosis are not always useful. However, clinical management of the disease is expected to benefit from the rapid pace of discoveries in the genomics of Alzheimer's.[xii] Most research on genome-based applications in Alzheimer's is still in early phases of the translational research process, which means that further research is still needed before their implementation can be considered. Considering the vast data implications of Alzheimer's and other related dementias, *Esperanza 3.0* is vital for several other reasons:

- Much of the care for Alzheimer's patients is working to maintain "moments of peace" and keep injuries (e.g., from falls) or infections (e.g., pneumonia) at bay. Where cures are often the desired end state for a disease, in the case of Alzheimer's, the care management plan is often focused on maintaining a patient's dignity, retaining cognitive function for as long as possible, and avoiding costly hospital admissions. Data is a powerful tool in navigating toward this goal and will help further this area of care greatly.
- Many neurological diseases, such as Alzheimer's, have a potential genetic component. And many adults with this disease in their family line remain in an apprehensive state. Rich longitudinal data that tracks these individuals as well as current patients will continue to inform what we do not know about Alzheimer's and related dementias. Since *Esperanza 3.0* provides a platform upon which to manage and monitor disparate

data streams (e.g., genetic, ecologic, environmental, epidemiological, clinical, laboratory, public health, geo-spatial), apply distributed and parallelized data processing across “big data” sets (e.g., whole genome sequences), and introduce business intelligence and data visualization techniques, facilitating high throughput analytics that allows data scientists to rapidly interrogate vast data sets and investigate diseases in novel ways.

- There is emerging evidence that epigenetic mechanisms, changes in organisms caused by modification of gene expression rather than alteration of the genetic code itself, contribute to Alzheimer's disease.[xiii] Epigenetic changes, whether protective, benign, or harmful, may help explain, for example, why one family member develops the disease and another does not. Within *Esperanza 3.0*, data scientists will have the opportunity to synthesize genetics, environmental and epidemiological information to learn more about Alzheimer's-related epigenetics, with the hope of developing individualized treatments based on epigenetic markers and their function.
- The number of Americans surviving into their 80s, 90s and beyond is expected to grow dramatically due to medical advances, as well as social and environmental conditions. Additionally, a large segment of the U.S. population, the baby boom generation, has begun to reach age 65 and older, when the risk for Alzheimer’s and other dementias is elevated. By 2025, the number of people age 65 and older with Alzheimer’s disease is estimated to reach 7.1 million -- almost a 40 percent increase from the 5.2 million age 65 and older affected in 2016. By 2030, the segment of the U.S. population age 65 and older will increase substantially, and the projected 74 million older Americans will make up over 20 percent of the total population. As the number of older Americans grows rapidly, so too will the numbers of new and existing cases of Alzheimer’s disease.[xiv]

- As the U.S. population ages, Alzheimer's is becoming a more common cause of death. Although deaths from other major causes have decreased significantly, official records indicate that deaths from Alzheimer's disease have increased significantly. Between 2000 and 2013, deaths attributed to Alzheimer's disease increased 71 percent, while those attributed to the number one cause of death (heart disease) decreased 14 percent. The increase in the number and proportion of death certificates listing Alzheimer's as the underlying cause of death reflects both changes in patterns of reporting deaths on death certificates over time as well as an increase in the actual number of deaths attributable to Alzheimer's. [xv]
- Improved surveillance of chronic diseases is essential for evidence-based advocacy and for raising political awareness and commitment; if the scale of the problem is invisible, as with the chronic disease burden, to argue in support for prioritization of chronic disease surveillance is difficult.

Research demonstrates that over eighty percent of brain disorders are associated with genomics defects in conjunction with environmental factors and epigenetic phenomena.[xvi] In 2016, Cacabelos, et al, have found that epigenetic modifications are related to disease development, environmental exposure, drug treatment and aging. Epigenetic changes are reversible and can be potentially targeted by pharmacological intervention. Both hypermethylation and hypomethylation of DNA, chomatin changes and miRNA dysregulation are common in age-related disorders and in many neuropsychiatric, neurodevelopmental and neurodegenerative disorders. Major epigenetic mechanisms may contribute to Alzheimer's disease (AD) pathology.[xvii]

The advent of high-throughput, next-generation sequencing (NGS) technology – rapid, accessible, cost-effective, whole genome sequencing (WGS) -- allows the identification of traceable differences in the pathogen genome using massively parallel processing. NGS facilitates the generation and detection of millions of long and short sequencing reads on a single machine run without the need for cloning – in orders of magnitude higher resolution as compared to earlier typing methods.[xviii]

Transforming the ability to interpret high resolution disease transmission, NGS reduces the processing cost and sequencing time down to a fraction of that of its predecessor, the Sanger chain-determination method, long-considered the gold standard for DNA sequencing[xix]. Adoption of NGS by public health laboratories has accelerated notably over the last decade, catapulted forward by infectious disease outbreaks with significant morbidity and mortality, e.g., the cholera epidemic in Haiti following the 2010 earthquake[xx] and E. coli 0104:H4 disease pandemic in 2010 associated with fenugreek sprout consumption[xxi]. A global implementation of standardized WGS in infectious disease surveillance is being advanced around the world and chronic disease surveillance will almost certainly follow suit.

Next-generation sequencing (NGS) is revolutionizing the speed and ability to trace the etymology and pathogenesis of disease by enabling the synthesis of multiple sequencing workflows seamlessly. For example, species, serotypes, virulence characteristics, and antibiotic resistance can be extracted from genomes and combined with phylogenetic, relevant subtyping information. [xxii]

Massively parallel processing employs short read molecular-based technologies to produce billions of nucleotide sequences during each run, and each genome is sequenced multiple times in small random pieces to generate very large datasets.[xxiii] There is a significant body of evidence that the confluence of next-generation sequencing (NGS), cloud computing, and advanced analytics are producing a paradigm shift in our understanding of disease etiology, development and treatment, and subsequently challenging the traditional surveillance model.[xxiv] [xxv]Advances in computing power and "big data" provide the engine for mathematical and statistical techniques by which disparate datasets can be synthesized and analyzed. Whole-genome sequencing offers precision resolution in orders of magnitude over earlier genotyping methods, transforming our approach to monitoring, controlling, and preventing diseases in both epidemic and endemic contexts. In addition, the increased availability of dense data characterizing the substrate population and the development of advanced computation and analytical tools to organize and interpret these large datasets broadens the potential for application of such data to high-resolution epidemiological problems.

TECHNICAL PROPOSAL

Esperanza 3.0 provides a platform upon which to manage and monitor data streams, apply distributed and parallelized data processing across “big data” sets (e.g., whole genome sequences), and introduce business intelligence and data visualization techniques. Federal and international bioinformatics partners will be able to securely access a wide range of disease data, risk factor indicators, and policy measures to describe the burden of Alzheimer’s and related dementias as well as common risk factors, identify research gaps, monitor population trends, and evaluate programs.

As *Esperanza 3.0* matures, advanced predictive and prescriptive analytics will offer data scientists increased opportunities to aid in evidence-based preventive efforts (e.g., early detection, disease prevention and intervention, and capacity planning). Advanced decision support from cognitive computing engines, cluster analysis, machine learning, and natural language processing can help providers recognize diagnoses that might remain elusive otherwise, while population health management tools can highlight those most at risk of being readmitted to the hospital or developing costly neurocognitive disorders and dementias.

DESIGN PROPOSAL

The objectives of *Esperanza 3.0* are to implement an enterprise data integration, management, analysis, and visualization platform utilizing best-in-breed technology to:

1. Integrate critical epidemiological, laboratory, environmental, situational awareness, countermeasure and response, and other known and *ad hoc* data collections and systems.
2. Build a platform that is:
    a. Designed with standardized processes;
    b. Able to facilitate stakeholder collaboration and real-time data sharing;
    c. Accessible through a single graphical user interface on mobile and PC platforms;
    d. In compliance with security requirements;
    e. Driven by analytic requirements;
    f. Scalable;
    g. Adaptable to all public health events; and
    h. Interoperable with existing IT systems.
3. Create a flexible and powerful analytic interface.
4. Visualize data from multiple data streams in a single unified view.
5. Develop user-friendly visualization tools to appropriately display data to support event response activities.
6. Support timely processing and provisioning of new data sources for platform users.
7. Securely share data with external partners.

LOGICAL ARCHITECTURE (LA)

Logical Architecture (LA) is the manner in which logical components of a solution are organized and integrated. Here is a depiction of the proposed logical architecture for *Esperanza 3.0 Microservices Architecture* (See Figure 3). Microservices architecture is a modern method of developing software applications as a suite of small, independently deployable, modular services in which each unit runs a unique process and communicates through a well-defined, lightweight mechanism to serve a discrete public health goal. As depicted in Table 1, the key features of microservices include: componentization as services, products not projects, organize around business capabilities, smart endpoints and dumb pipes, decentralized governance and data management, infrastructure automation, designing for failure (fault tolerance), and evolutionary design. We will employ containerization, the process of developing independently deployable, replaceable, upgradable services. Introducing container technology, such as Docker, can be employed to streamline the development and testing process. Docker is an open source platform for packaging, distributing, and managing apps within containers. The isolation between containers running on the same host simplifies deploying microservice code developed using different languages and frameworks. By allowing more containers in the environment, we reduce the need to provision additional servers as containerization increases scalability anywhere from 10 to 100 times that of traditional virtual environments.[xxvi]

Figure 3. *Esperanza 3.0* Microservices Architecture

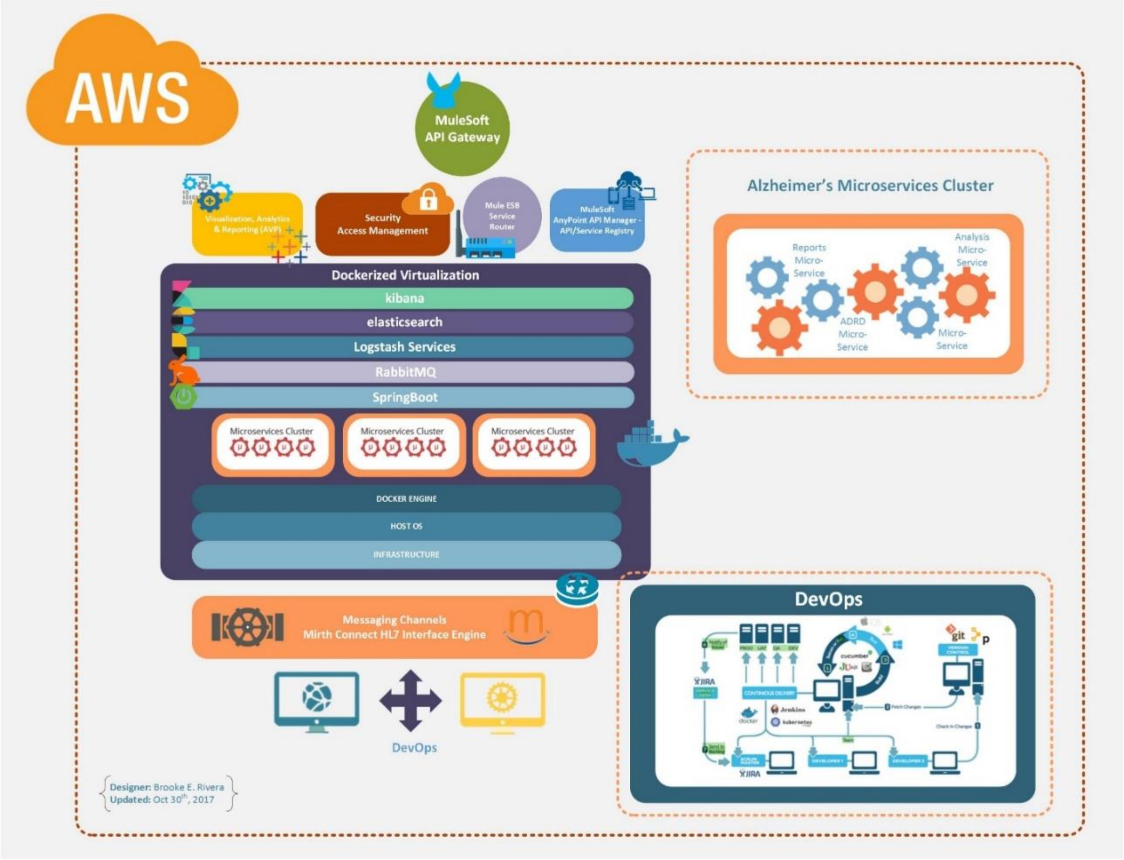


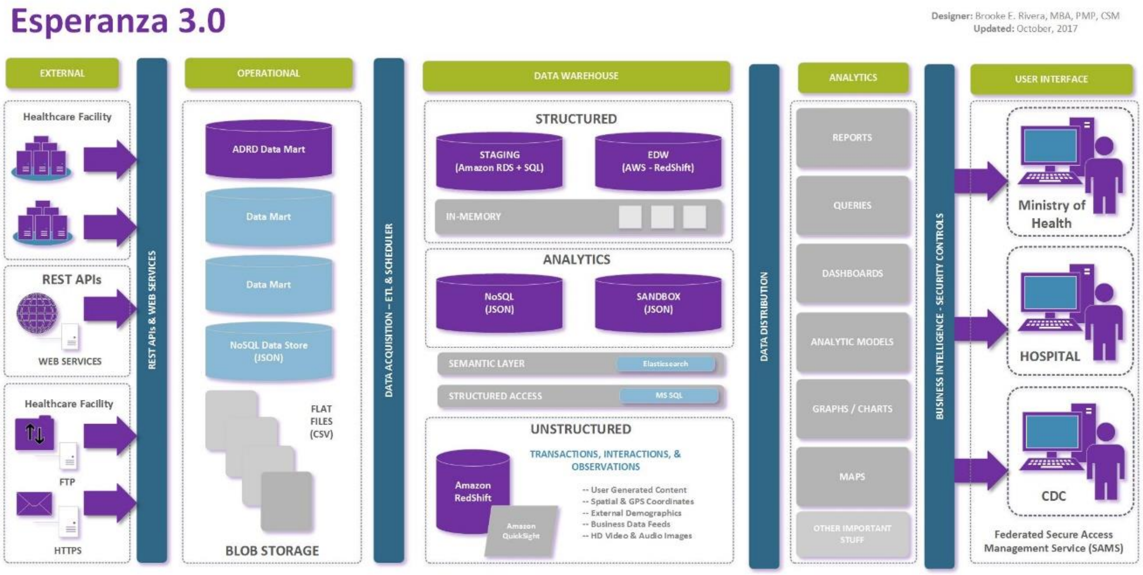
Table 1. Key Features of a Microservice Architecture

Key Feature	Description
Componentization as services	Dividing large services into small sets of services which communicate with each other independently without going through a common messaging layer.
Products not Projects	The product mentality, ties in with the linkage to business capabilities. Rather than looking at the software as a set of functionalities to be completed, there is an on-going relationship where the question is how can software assist its users to enhance the business capability.
Organize around business capabilities	Organize around business capabilities
Smart endpoints and dumb pipes	Applications built from microservices aim to be as decoupled and as cohesive as possible - they own their own domain logic and act more as filters in the classical sense - receiving a request, applying logic as appropriate and producing a response.
Decentralized governance	Use different types of data stores to store different data instead of having a centralized one
Decentralized data management	Use different types of data stores to store different data sets instead of managing a single, centralized repository.
Infrastructure automation	Expedited configuration management and deployment; quickly spin up new instances based on runtime features.
Design for failure	A consequence of using services as components is that applications need to be designed so that they can tolerate the failure of services, and they must fail as gracefully as possible.
Evolutionary design	Incremental development using a concept like MVP (minimum viable product) and improving the product iteratively over time.

ENTERPRISE ARCHITECTURE (EA)

The proposed Enterprise Architecture (EA), a conceptual blueprint that defines the structure and operation of an organization, in this case a cloud-based data warehouse, is depicted in Figure 4. A data warehouse is different from an operational database in that it is built to facilitate the analysis of historical data as opposed to handling transactions. This means that data warehouses tend to be orders of magnitude larger than their corresponding operational databases. This data warehouse will allow data scientists to run complex analytic queries against petabytes of structured data, using sophisticated query optimization, columnar storage on high-performance local disks, and massively parallel query execution. The cluster scales as performance and capacity needs change. All data warehouses are meant to handle large volumes of data, but this cloud-based data warehouse will be optimized for clusters of different sizes. Notably, in cloud-based warehouses, a major consideration is how quickly and straight forward it is to provision new resources, and whether computing and storage resources can be provisioned separately.

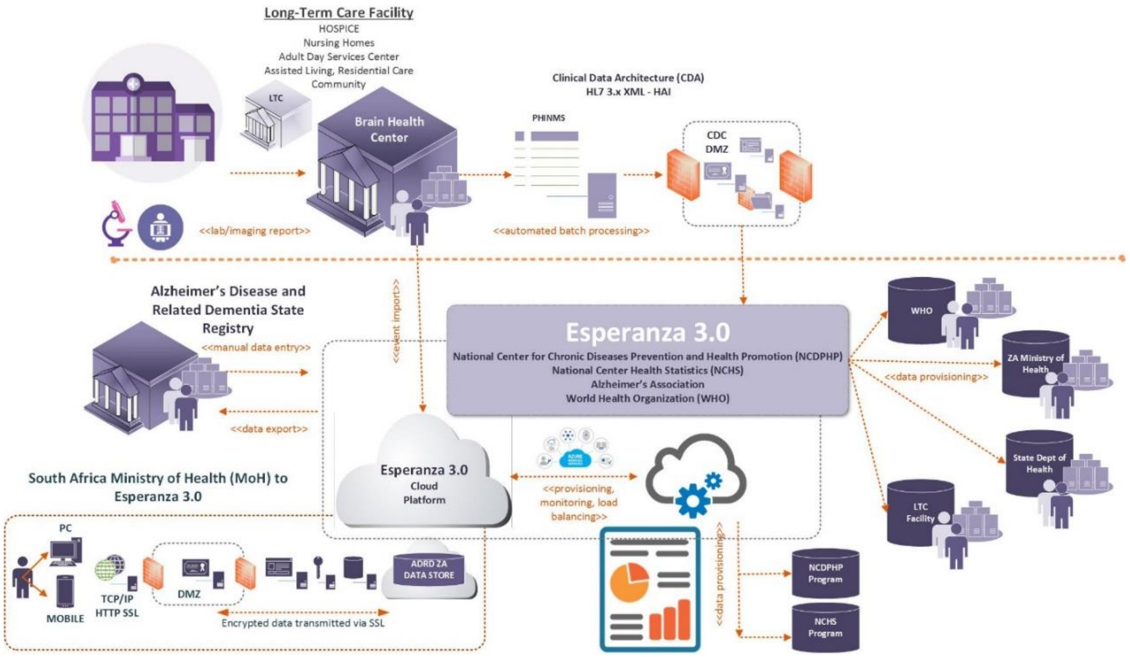
Figure 4. *Esperanza 3.0* Enterprise Architecture (EA)



INFORMATION FLOW DIAGRAM (IFD)

A data flow diagram is a graphical representation of the "flow" of data through an information system, modelling its process aspects. In the future, healthcare partners (eg., long term care facilities, dementia registries, laboratories, imaging centers) from participating nation-states will transmit electronically, or upload directly via the *Esperanza 3.0* user interface, case notifications, imaging records, and laboratory reports. This data will be ingested, stored, transformed, and synthesized for stakeholders at the CDC, WHO, and other authenticated users for the purposes of advanced analytics, visualization, and reporting -- allowing data scientists to rapidly interrogate vast data sets and investigate dementia in novel ways. They will be able to access massively parallelized data sets to conduct comparative analysis of structural variants and rearrangements in human and malignant genomes, with emphasis on data integration and uncertainty visualization, conduct epigenomic data clustering, run predictive and prescriptive algorithms, and to discover, annotate and easily share research observations on this collaborative platform, *Esperanza 3.0*.

Figure 5. *Esperanza 3.0* Information Flow Diagram


Long-Term Care Facility
HOSPICE
Nursing Homes
Adult Day Services Center
Assisted Living, Residential Care
Community
LTC
Brain Health Center
<<lab/imaging report>>
Clinical Data Architecture (CDA)
HL7 3.x XML - HAI
PHINMS
<<automated batch processing>>
CDC
DMZ
Alzheimer's Disease and Related Dementia State Registry
<<manual data entry>>
<<data export>>
<<event import>>
Esperanza 3.0
National Center for Chronic Diseases Prevention and Health Promotion (NCDPHP)
National Center Health Statistics (NCHS)
Alzheimer's Association
World Health Organization (WHO)
WHO
ZA Ministry of Health
<<data provisioning>>
State Dept of Health
LTC Facility
South Africa Ministry of Health (MoH) to Esperanza 3.0
Esperanza 3.0
Cloud
Platform
<<provisioning, monitoring, load balancing>>
PC
MOBILE
TCP/IP
HTTP SSL
DMZ
ADRD ZA
DATA STORE
Encrypted data transmitted via SSL
<<data provisioning>>
NCDPHP Program
NCHS Program

SUMMARY

As we face a new decade in the fight against Alzheimer's, the reality is that so much about this disease is still unknown and there are more questions than answers. What drives disease progression? What treatments are most effective? How can we help afflicted families? After reviewing research findings over the last two decades, initiatives and information on treatment and prevention, I was heartened by what I found. There is hope. *Hay esperanza.*

[i] Cacabelos R (2016) *Epigenetics of Brain Disorders: The Paradigm of Alzheimer's Disease*. J Alzheimer's Dis Parkinsonism 6:229. doi: 10.4172/2161-0460.1000229

[ii] Tripathi, R., et al. (2016, April). *Next-generation sequencing revolution through big data analytics.* Taylor and Francis Online. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/21553769.2016.1178180

[iii] Thakur, R. S., et al. (2012, Dec). *Now and Next-Generation Sequencing Techniques: Future of Sequence Analysis Using Cloud Computing.* Frontiers in Genetics. U.S. National Library of Medicine, National Institutes of Health. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3518790/

[iv] Giteria, O., et al. (2014, May). *Implementation of Cloud-based Next Generation Sequencing data analysis in a clinical laboratory.* BioMed Central. U.S. National Library of Medicine, National Institutes of Health. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4036707/

[v] He, K. Y., et al. (2017, Feb). *Big Data Analytics for Genomic Medicine.* U.S. National Library of Medicine, National Institutes of Health. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5343946/

[vi] Kashyap, H., et al. (2014, Sept). *Big Data Analytics in Bioinformatics: A Machine Learning Perspective.* Retrieved from https://arxiv.org/pdf/1506.05101.pdf

[vii] Luo, J, et al. (2016, Jan). *Big Data Application in Biomedical Research and Health Care: A Literature Review*. Biomedical Informatics Insights. U.S. National Library of Medicine, National Institutes of Health. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4720168/

[viii] Jenet, J Appl. (2011, Nov). *Sequencing and genome sequencing*. Journal of Applied Genetics. U.S. National Library of Medicine, National Institutes of Health. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3189340/

[ix] Huang, X., et al (2009, June). *High-throughput genotyping by whole-genome resequencing.* Genome Research. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2694477/

[x] Harry, J. (2013, March) Alzheimer's Association. *Testimony of Harry Johns, President and CEO of the Alzheimer's Association Fiscal Year 2014 Appropriations for Alzheimer's-related Activities at the U.S. Department of Health and Human Services*. Public Policy Office. Retrieved from https://www.alz.org/documents/national/submitted-testimony-050113.pdf

[xi] Dementias, including Alzheimer's Disease. HealthyPeople.gov. Office of Disease Prevention and Health Promotion (ODPHP). Retrieved from https://www.healthypeople.gov/2020/topics-objectives/topic/dementias-including-alzheimers-disease

[xii] Mihaescu, R., et al. (2010). *Translational research in genomics of Alzheimer's disease: a review of current practice and future perspectives.* Journal of Alzheimer's Disease. PubMed.gov. U.S. National Library of Medicine, National Institutes of Health. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20182048

[xiii] Mastroeni, et al. (2011, July). *Epigenetics Mechanisms in Alzheimer's Disease.* Neurobiological Aging. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3115415/

[xiv] Gaugler, J., et al. (2017). Alzheimer's Association Report: 2017 Alzheimer's disease facts and figures. Alzheimer's Association. Retrieved from http://www.alzheimersanddementia.com/article/S1552-5260(17)30051-1/pdf

[xv] Alzheimer's.org (2017). *2017 Alzheimer's Disease Facts and Figures.* Alzheimer's Association. Retrieved from https://www.alz.org/facts/

[xvi] Yegambaram, M., et al. (2015, Feb). *Role of Environmental Contaminants in the Etiology of Alzheimer's Disease: A Review.* Current Alzheimer Research. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4428475/

[xvii] Cacabelos, R. et al. (2016, April). *Epigenetics of Brain Disorders: The Paradigm of Alzheimer's Disease.* Journal of Alzheimer's Disease and Parkinsonism. Retrieved from https://www.omicsonline.org/open-access/epigenetics-of-brain-disorders-the-paradigm-of-alzheimer-s-disease-2161-0460-1000229.php?aid=72455

[xviii] Reuter, J, et al. (2015, May). *High-Throughput Sequencing Technologies.* U.S. National Library of Medicine, National Institutes of Health. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494749/

[xix] James, H. and Chain, B. (2016, Jan). Sequence of Sequencers. The History of Sequencing DNA. Genomics. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4727787/

[xx] Heitman, J. (2014, April). *The 2010 Cholera Outbreak in Haiti: How Science Solved a Controversy.* U.S. National Library of Medicine, National Institutes of Health. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3974815/

[xxi] MMWR Report. (2013, De). *Outbreak of Escherichia coli O104:H4 infections associated with sprout consumption - Europe and North America, May-July 2011. Pubmed.gov.* Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/24352067

[xxii] Pak, T., Kasarskis, A. ((2015, Dec). *How Next-Generation Sequencing and Multiscale Data Analysis Will Transform Infectious Disease Management.* Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4643486/

[xxiii] Nadon, C., et al. (2017, June) *PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance.* U.S. National Library of Medicine, National Institutes of Health. Eurosurveillance. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5479977/

[xxiv] Tripathi, R., et al. (2016, April). *Next-generation sequencing revolution through big data analytics.* Taylor and Francis Online. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/21553769.2016.1178180

[xxv] Thakur, R. S., et al. (2012, Dec). *Now and Next-Generation Sequencing Techniques: Future of Sequence Analysis Using Cloud Computing.* Frontiers in Genetics. U.S. National Library of Medicine, National Institutes of Health. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3518790/

[xxv] Giteria, O., et al. (2014, May). *Implementation of Cloud-based Next Generation Sequencing data analysis in a clinical laboratory.* BioMed Central. U.S. National Library of Medicine, National Institutes of Health. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4036707/

[xxv] He, K. Y., et al. (2017, Feb). *Big Data Analytics for Genomic Medicine.* U.S. National Library of Medicine, National Institutes of Health. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5343946/

[xxv] Kashyap, H., et al. (2014, Sept). *Big Data Analytics in Bioinformatics: A Machine Learning Perspective.* Retrieved from https://arxiv.org/pdf/1506.05101.pdf

[xxvi] Chavis B, Jones T, et al. (2015, April). *Docker on AWS: Running Containers in the Cloud*. Retrieved from https://d0.awsstatic.com/whitepapers/docker-on-aws.pdf