Chapter 4
Data Platform for the Research and Prevention of Alzheimer's Disease

Ning An, Liuqi Jin, Jiaoyun Yang, Yue Yin, Siyuan Jiang, Bo Jing, and Rhoda Au

Abstract With the rapid increase in global aging, Alzheimer's disease has become a major burden in both social and economic costs. Substantial resources have been devoted to researching this disease, and rich multimodal data resources have been generated. In this chapter, we discuss an ongoing effort to build a data platform to harness these data to help research and prevention of Alzheimer's disease. We will detail this data platform in terms of its architecture, its data integration strategy, and its data services. Then, we will consider how to leverage this data platform to accelerate risk factor identification and pathogenesis study with its data analytics capability. This chapter will provide a concrete pathway for developing a data platform for studying and preventing insidious onset chronic diseases in this data era.

Keywords Data platform • Alzheimer's disease • Implementation

4.1 Background

In this chapter, we focus on the data platform that could be used to advance the research and prevention of Alzheimer's disease. At the beginning, some background about Alzheimer's disease will be introduced, including its social affects, biomarkers, epidemiological investigation, etc.

4.1.1 Alzheimer's Disease

Alzheimer's disease (AD) is a chronic neurodegenerative disease that usually starts slowly and gets worse over time, and it causes about 60–70% of cases of dementia.

N. An • L. Jin • J. Yang (✉) • Y. Yin • S. Jiang • B. Jing
School of Computer and Information, Hefei University of Technology, 193 Tunxi Road, Hefei, Anhui 230009, China
e-mail: jiaoyun@hfut.edu.cn

R. Au
School of Medicine, Boston University, 72 E. Concord Street, Boston, MA 02118, USA

B. Shen (ed.), *Healthcare and Big Data Management*, Advances in Experimental Medicine and Biology 1028, DOI 10.1007/978-981-10-6041-0_4

As medical science continues to reduce mortality, there is a concomitant increase in the risk for Alzheimer's disease (AD) and other types of dementia. A projected 13.8 million people in the United States are estimated to be diagnosed with AD by the year 2050 [1]. Despite this staggering number, the larger, more stunning global impact of this disease may well come from Third World countries where the health advances coupled with population growth strategies are converging to create a disproportionately large elderly population. In the year 2015, there were about 46.8 million people with dementia worldwide, and estimates are for 131.5 million by 2050, the majority of which will be diagnosed with Alzheimer's disease [37].

Besides, as a social disease, AD brings great burden to society and patients' family. From 2000 to 2010, the cost of caring for those with dementia/AD rose from $18 billion to over $600 billion. The burden of caring for people with AD is estimated to be $1.2 trillion by 2050 [2, 3]. It has become a serious disease affecting the global public health and attracted great attentions from global researchers. Numerous works have been done on this, for example, there are more than 112,000 research papers in PubMed up to 2015. Unfortunately, to date, there are still no disease-modifying medications. In order to tackle AD, the Obama administration and the National Institutes of Health have homed in on Alzheimer's disease, setting an ambitious goal to have an effective treatment for the brain-wasting disease by 2025.

Actually, some remarkable findings have been revealed, such as some biomarkers to determine AD's pathogenesis. Besides, research indicates that delaying onset by just 5 years will cut an individual's risk for diagnosis in half [4]. Identifying dementia risk factors that are amenable to intervention prior to disease onset represents a viable strategy to mitigate late-life dementia. And a number of dementia risk factor profiles have been developed from epidemiological studies. In the next two sections, we will introduce the biomarker findings and epidemiological investigations for AD.

4.1.2 Biomarkers for Alzheimer's Disease

A fundamental task for AD's research is to reveal the pathogenesis. Although the pathogenesis is still not clear, some biomarkers have been identified to help diagnose the early onset of the disease. This could guide the family to take some procedures to delay AD's deteriorating and help to reduce the cost of patients' healthcare.

Biomarkers could be extracted from various materials, such as blood, cerebrospinal fluid, MRI, PETs, genomes, etc. Current biomarkers that have been widely recognized include Aβ, Tau, APOE, etc.

- Abnormal amyloid-β-protein(Aβ) aggregation

  Researchers injected Aβ to the brains of transgenic mice, and found necrosis in the injected sites, missing in the surrounding neurons and glial proliferation [34]. Besides, these findings are closely related to the injected dose. After culturing human brain cortex neurons and Aβ together, some Aβ formed amyloid-β-protein aggregation which would lead to neuronal degeneration. It's quite similar between neuronal degeneration by Aβ and pathologic changing of Alzheimer's disease.

- Hyper-phosphorylated tau protein aggregation

  The total amount of tau protein in the normal is less than that in AD's brains, whose normal tau protein decrease and hyper-phosphorylated tau protein increase largely. Hyper-phosphorylated tau in AD's brains loses not only functions of promoting microtubule assembly formation but also the ability of the adhesion strength of the microtubule protein, decreased by 90%. Compared with abnormal amyloid-β-protein (Aβ) aggregation, recent studies have shown that neurofibrillary tangles (NFTs) caused by hyper-phosphorylated tau protein have a higher correlation with AD [34].
- Apolipoprotein E (APOE) ε4

  Early in 1994, several studies have proved APOE ε4 has a meaningful relationship with praecox AD, and familial AD patients had a higher gene frequency of APOE ε4 than the nonfamilial patients. Comprehensive studies reveal the correlation of APOE ε4 and AD, which is that age of onset would bring forward because of the high amount of APOE ε4. For example, people who were homozygotes for APOE ε4 would have an earlier age of onset than people who were heterozygote for APOE ε4. Naturally, people with no APOE ε4 would have a later age of onset. Therefore, it can be recognized that APOE ε4 was one of the most important determinants of AD [35].

4.1.3 Epidemiological Investigation for Alzheimer's Disease

While biomarkers' research aims to determine the pathogenesis and guide the treatment direction, prevention through modifiable risk factors remains the most cost-effective strategy for combatting the disease as evidenced by declining AD risk despite the lack of effective treatment options. Hence, various epidemiological investigations have been done to find AD-related risk factors.

Lifestyle risk factors including physical activity, sleep, and diet also impact risk for cognitive decline and AD. To a more limited extent, environmental exposures have also been linked to brain health. An important category of risk factors is cardiovascular risk factors (CVRF). Numerous studies suggest CVRF are associated with increased risk for cognitive decline and AD. Zlokovic's "two-hit vascular hypothesis of AD" states that vascular risk leads to neuronal injury and dysfunction (hit one), suggesting an initial non-amyloid pathway [9]. Identifying the specific role of potentially modifiable CVRF have on long-term cognitive health could lead to interventions that will allow lengthening years of active life and life quality. Delaying onset of CVRF by just 5 years cuts a person's risk by 50%.

Risk factors' discovery relies on epidemiological investigations by developing various protocols. Framingham Heart Study (FHS) is a pioneer in this area [5]. It was first initiated in 1948, and embarked on a study of heart disease, and is credited with identifying many of the known risk factors for cardiovascular disease. FHS has since applied its well-developed longitudinal study design to the investigation of prevalent and incident dementia and associated risk factors. The 2015 Alzheimer's Association estimates of lifetime risk of dementia and Alzheimer's disease utilized data from FHS' dementia study [6]. Some successful protocols they applied include socioeconomic, medical history questionnaire, blood and urine, physical activity, sleep questionnaire, food frequency questionnaire, environmental exposure, cardiovascular assessment, depression assessment, cognitive assessment, etc.

Traditional strategy for epidemiological investigation is usually based on questionnaire, and the investigation is usually done by face to face. However, more recent integration of wearable technology has allowed much more extensive assessment of physical activity and heart rate linked to cognitive impairment and brain structure, such as the Shimmer sensing platform could provide more complete, more continuous, and more accessible dynamics of the physiological signals associated with various levels and types of physical activity [7]. Besides, some digital devices could provide more accuracy assessment, e.g., the digital pen in place of a regular ballpoint pen for participant drawn tests [8]. The digital pen allows collection of decision-making latencies and graphomotor characteristics that may reflect subtle differences in underlying cognitive processing.

4.2 Public Data Platform

The biomarkers' research and epidemiological investigation have generated a large amount of data. These data comes from various institutes, and usually each institute could only make use of its own data. This limited data access may restrain the advance of this area on some levels, since some underlying knowledge could only be achieved when the data size reach a finite number. Therefore, some organizations turn to construct a uniform data platform to integrate these data from various institutes such that these numerous amounts of data could be fully utilized.

4.2.1 Public Data Platform for Alzheimer's Disease

- NACC (https://www.alz.washington.edu/)

  NACC is a public data platform established by the National Alzheimer's Coordinating Center. It provides services for the investigators of National Alzheimer's Centers, the officers of National Institute on Aging, staff of National Alzheimer's Coordinating Center, as well as the public [10].

  Data contained: NACC consists of three data sets, including Minimum Data Set (MDS), Neuropathology Data Set (NP), and Uniform Data Set (UDS). These data sets cover data from questionnaires, medical images, genetics, and clinical examinations.

Service provided: data quality control, data share, data report generation, and data directory

- ADNI (http://adni.loni.usc.edu/)

  ADNI is built by Alzheimer's Disease Neuroimaging Initiative. Investigators could use the data to define each phrase of AD and to make prediction to AD. The data comes from ANDI's study participants, including Alzheimer's disease patients, mild cognitive impairment subjects, and the elderly [11, 12].

  Data contained: ADNI contains four types of data, including (1) clinical data, such as the recruitment information of each subject, demographic information, physical information, recognition assessment, CSF concentration, and the radiative data of biomarker; (2) MRI and PET image data, which could be used to track both the progression of AD and changes in the underlying pathology; (3) genetic data, such as genotyping and sequencing data; and (4) biospecimen data, such as blood, urine, and cerebrospinal fluid (CSF) data from participants.

  Service provided: data share, expert support, and data usage instruction
- AMP-AD Knowledge Portal (https://www.synapse.org)

  The platform is the knowledge portal for accelerating medicine partnership, AD target discovery, and preclinical validation. The main aim is to shorten the time between new medicine development and AD prevention. The platform provides a community for various consortiums to share data, data analysis method, and computing model [13].

  Data contained: RNA data, DNA data, genome data, protein data, and MRI data, among these data, some from human samples and some from animal samples

  Service provided: data usage agreement, data share, data analysis results share, data directory, besides, the platform provide query language to query the data; R, python, or command line could be applied to download the data file.
- The AddNeuroMed Study

  AddNeuroMed is funded by the EU FP6 program and is a public-private partnership for AD's biomarker discovery and replication. The platform aims to help the organization to share the data and integrate resources [14, 15].

  Data contained: (1) clinical data set, 700 samples; (2) DNA data set, collected from blood, 644 samples; (3) brain image data, 175 samples; (4) protein data, collected from blood, 645 samples; and (5) mRNA data, collected from blood, 674 samples

  Service provided: data usage agreement, data share, data query, and data analysis results share
- NIAGADS (https://www.niagads.org)

  National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) is developed by NIA as a genetic data repository, which aims to share data for investigators studying the genetics of late-onset AD. The data is mainly used for investigators to do secondary research after the publication of any patent or research [16].

Data contained: genetic data generated from the analysis of patients' cerebrospinal fluid (patients' CHR, SNP), clinical examination data, and investigation data

Service provided: data share, data query, data comparison and annotation, web tools for analyzing large-scale genome data, and various software interfaces

- The Johns Hopkins Alzheimer's Disease Research Center

  ADRC is funded by NIA and established in 1984. The organization has made great contribution to the symptom and pathology research of AD. It aims to find AD's treatment in pathology. They provide both mental and physical ways to cure the patients [17].

  Data contained: blood and DNA data, brain specimen data, clinical data, and cognitive assessment data

  Service provided: data share and introduction to clinical research
- Alzheimer's Disease Cooperative Study

  ADCS is a cooperative organization between NIA and the University of California. It is the major Alzheimer's disease clinical research institute of the federal government, which aims to treat Alzheimer's disease for its physical and cognitive symptom [18].

  Data contained: cognitive assessment data and biomarkers' data

  Service provided: data share, integrated analysis tools to measure cognitive ability, life function, physical information, and quality of life

4.2.2 Classical Health-Related Data Platform

- CHARLS (http://charls.ccer.edu.cn)

  China Health and Retirement Longitudinal Study (CHARLS) aims to analyze Chinese aging problem and promote the interdisciplinary research based on the collection of a set of micro-high-quality data from typical families or persons older than 45 in China. The CHARLS baseline research began in 2011. It covers 150 counties and 450 villages and contains about 17,000 people from 10,000 families [19].

  Data contained: CHARLS consists of 12 data sets, including demographic background; family information; family transfer; health status and functioning; healthcare and insurance; work, retirement, and pension; household income; individual income; housing characteristics; interviewer observation; weights; and PSU.

  Service Provided: data share and data research results share
- cBioPortal

  The platform was first developed by the Memorial Sloan Kettering Cancer Center and now developed and maintained by multiple institutions including MSK, the Dana Farber Cancer Institute, Princess Margaret Cancer Centre in Toronto, Children's Hospital of Philadelphia, The Hyve in the Netherlands, and

Bilkent University in Ankara, Turkey. The platform aims to provide the visualization, analysis, and download of large-scale cancer genetic data [20].

Data contained: DNA copy-number data, mRNA, microRNA expression data, non-synonymous mutations, protein-level and phosphoprotein-level data, DNA methylation data, and limited de-identified clinical data

Service provided: data query, data share, webAPI (help access the txt or uml format data via uri), develop kit of R and MATLAB, visualization tools, and visualization tutorial

- CommonMind Consortium Knowledge Portal

  CMS Knowledge Portal aims to share the data and data analysis generated by CommonMind organization in a transparent, reproductive way. The platform commits to share the data, analysis results, and the source codes [21].

  Data contained: (1) DNA data set from schizophrenia, bipolar disorder, and mood disorder patients, 621 samples in total; (2) clinical data set that contains clinical data and metadata, 621 samples in total; and (3) RNA sequence data set from schizophrenia, bipolar disorder, and mood disorder patients, 613 samples in total

  Service provided: data share, data query, data analysis result share, and introduction to data analysis method

- US CDC Healthy Brain Initiative

  The platform is based on the healthy aging project of CDC, with the purpose of promoting the health and life quality of elder people. The initiative aims to integrate the public health and aging service network to promote the prevention of elder people's disease and promote the health-related life quality [22, 23].

  Data contained: general health data, nutrition, physical function data, vaccine relevant data, alcohol and smoking data, mental health data, and cognitive health data

  Service provided: data share, data visualization, API for data access, health education for the aged, and disease prevention

- NCMI (http://www.ncmi.cn)

  The National Scientific Data Sharing Platform for Population and Health (NCMI) serves for technology innovation, government administration, and development of medical and health service and provides data share service for cultivation of innovative talents and development of public health industry [24].

  Data contained: fundamental medical data, clinical medicine data, public health data, traditional Chinese medicine data, pharmacy data, population and reproductive health data

  Service provided: data share, data query, disease prevention, standard document, and data analysis tools (sequence and structure to predict mRNA)

- The Health Indicators Warehouse (http://www.healthindicators.gov/)

  The Health Indicators Warehouse (HIW) aims to use the high-quality data to help people understand the health status and determinants of community and to facilitate the prioritization of interventions. It could meet the need of various population health initiatives and provide a single, user-friendly data source for national, state, and community [25].

Data contained: health indicators including demographic, food health, situation of chronic disease, etc. and 1291 indicators in total

Service provided: data directory, data share, WEBAPI, and web service such as RESTFUL, SOAP, and EDGE

- The Michael J Fox for Parkinson (https://www.michaeljfox.org/)

  The platform is founded by Michael J. Fox Foundation and is dedicated to find strategies to cure the Parkinson disease via funded studies and ensured development of improved therapies [26].

  Data contained: DNA, RNA, CSF, and some biospecimens

  Service provided: data share, disease education, and biospecimen share

4.3 Data Platform for Alzheimer's Disease

4.3.1 Objective of Data Platform

Compared with other health-related data platforms, current AD data platform mostly focuses on the data collection and access, and only a few services could be achieved by researchers. In order to facilitate the AD research and prevention, an idle AD data platform should satisfy the following objectives:

- Integrate various institutes' data, and supply authorized data access to enable the big data analysis for AD.
- Provide comprehensive data services to simple researchers' analysis, such as data management, data visualization, data statistics, data analysis tools, etc.
- Offer application programming interface (API) to promote AD's prevention.

According to the above objectives, we aim to construct a data platform framework for AD and specify its detailed services and technologies.

4.3.2 Framework of Data Platform

The data platform consists of four modules, including entrance module, service module, data module, and computing module, which are demonstrated in Fig. 4.1. Users collect various types of data and upload them through entrance module under the agreement. Users that want to access the data should make a request and get the service under the permission of the committee. The functions of the four modules are described below:

- Entrance Module

  Entrance module is used to achieve the load balance and reverse proxy of various requests. It commutes with service module to orient users' http request to

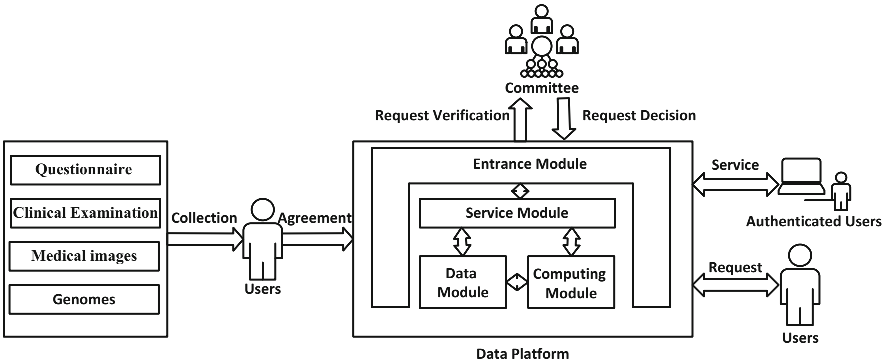

Fig. 4.1 Framework of data platform. The platform consists of four modules: entrance module, service module, data module, and computing module. Entrance module receives various requests and commutes with service module to orient these requests to the corresponding service of a specific node in the server. Service module calls functions in data module and computes the module to accomplish the requests. Data module is responsible for data storage, and computing module handles the computation tasks

the corresponding service of a specific node in the server. Besides, it also identifies users' authority.

- Service Module

  Service module receives users' requests from entrance module and handles these requests by calling the corresponding service function in this module. Sometimes, it should also refer to the functions in data module and computing module to realize the service function. Thus, commutations would happen between service module and data module or computing module.

- Data Module

  Data module is responsible for organizing various types of data. It provides APIs for service module and computing module, such that these two modules could achieve data access. It should also balance the storage load.

- Computing Module

  Computing module is in charge of various computing operation. It integrates various computation functions, such as statistics, machine learning modeling, etc. Hence it needs to communicate with data module for data access and with service module for returning data computation results. The computation of load balance is also realized in this module.

4.4 Services of Data Platform

The data platform fulfills its objectives by various publicly available services, including data directory, data share, data management, data analysis support, etc. Figure 4.2 illustrates the composition of platform's services, and the detailed descriptions of these services will be introduce below.

4.4.1 Data Directory Service

Data directory service aims to provide users the landscape of data stored in the data platform, such as data types, data size, etc. Besides, users could query these data information. The following data types should be contained in the data platform.

- Questionnaire data

  This type of data consists of various epidemiological investigation data, including socioeconomic, medical history, physical activity, and sleep questionnaire, food frequency questionnaire, environmental exposure, cardiovascular assessment, depression assessment, cognitive assessment, etc. This kind of data usually takes numeric or text format. By analyzing these data, risk factors could be determined to identify AD's prevention strategy or achieve AD screening.
- Clinical examination data

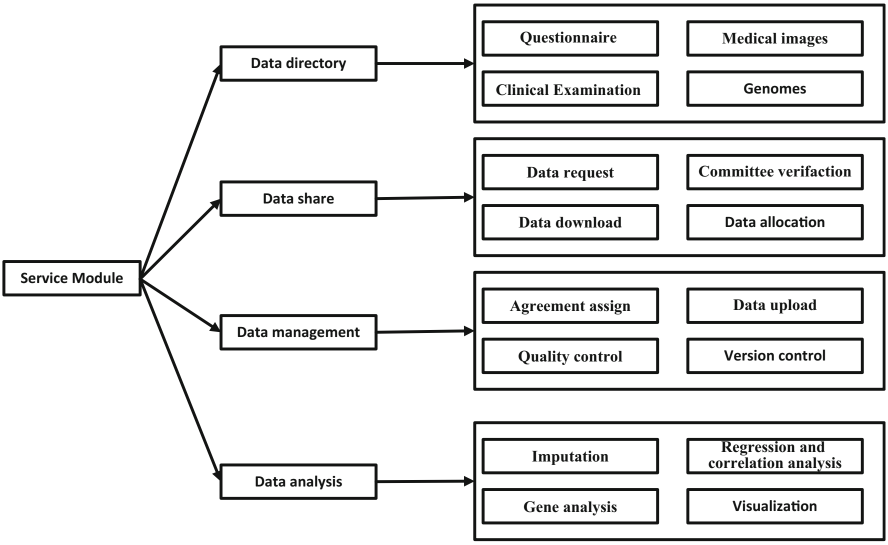

Fig. 4.2 Services provided by the data platform. The service module contains four types of service, including data directory, data share, data management, and data analysis support

This type of data is obtained through various clinical examinations, including blood examination, urine examination, physical examination, etc. The format of these data is the same as questionnaire data, i.e., numeric or text. Biomarkers and risk factors could be derived from these examination data to help reveal AD's pathogenesis, identify AD's prevention strategy, or diagnose the disease.

- Medical images' data

  Medical images' data comprises of various images, such as magnetic resonance imaging (MRI), positron emission tomography (PET), electrocardiogram (ECG), etc. These data are usually used to derive biomarkers to help diagnose AD.
- Genomes' data

  Genomes' data is the genomes' information for each individual. These could be represented as sequences in database. Researchers analyze genomes to find biomarkers in order to reveal AD's pathogenesis.

4.4.2 Data Share Service

An important reason of integrating data from various institutes is to break the data access barrier and make big data available for researchers. Hence, data share service is a fundamental service.

Data are collected from different institutes, and privacies need to be protected. Thus, data access should obey an authorized procedure. First, researchers upload their access request, including types, sizes, plan of utilization, etc. Then the request will be transferred to the community for verification. Once the request gets improvement, a download link will be sent to the researchers for access. Besides, an agreement needs to be assigned by the requesters. The agreement indicates privacy protection, data share protocol, limit of data utilization, etc.

4.4.3 Data Management Service

Researchers usually collect their own data for analysis; therefore, they need to organize and manage these data, which is complicated and time-consuming. In order to simplify these processes, the platform provides a data management service for researchers to upload their data and conduct management.

The first concern needing to care about is the protection of each researcher' data, e.g., who can access these data, etc. The protection could be classified into four levels, including private, limited access controlled by the owner, limited access controlled by the platform community, and public. These levels indicate the data could only be accessed by the owner, accessed with the owner's permission, accessed with the platform community's permission, and accessed publicly, respectively. Before a researcher uses the platform to manage its data, an agreement needs to be assigned to illustrate the protection level, each responsibility, etc.

The data platform provides public interface for users to upload their data, and data directory service in Sect. 1.4.1 defines the data types. Besides, the interface should clarify data format for data storage. There have been some successful format including csv, plink, vcf, fasta, etc. They could be adopted by the data platform.

After uploading the data, users could revise the pretension level of their data and utilize platform's service to achieve data quality control, data analysis, etc. Quality control consists of data imputation, data quality verification, data collection process control, etc. Data analysis refers to generating data report by visualization, statistics analysis, data mining, etc.

Besides, the platform provides version control function for users' data. The user could choose to roll data, and create data labels, such that users could manage numerous amounts of data with frequent change, and the platform could still guarantee data's safety.

4.4.4 Data Analysis Support

Data analysis is a necessary strategy for researchers to achieve inspiring results. The platform offers comprehensive data analysis functions to support researchers' analyzing requirement. These functions include data imputation, machine learning modeling, genome analysis, and alignment, which are described below:

- Data imputation

  There may be missing values in the collected data due to various reasons, e.g., misoperation, noncooperation, etc. If analyzing the data with missing values directly, some bias may be introduced, resulting in unreliable outcomes. Imputing these missing values could make up these disadvantages to some extent. Therefore, the platform adopts some classical imputing methods, such as expectation-maximization (EM), k-nearest neighbors (KNN), least squares fitting (LLS), etc. Users could choose these imputation options to fill up missing values.
- Machine learning modeling

  Sometimes, researchers need to determine the relationships between various types of data or apply these data to make prediction. Then machine learning models should be established for these analyses. The platform incorporates diverse machine learning modeling strategies to fulfill this requirement, for example, correlation analysis for identifying relationships; logistic regression, SVM, or neural networks for predicting specific values; k-means or spectral clustering for unlabeled data clustering; principal component analysis (PCA) or locally linear embedding (LLE) for dimensionality reduction; etc.
- Genome analysis

Biomarkers could be derived from genomes; thus, genome analysis may be needed to achieve this function. The platform would integrate some basic genome analysis tools, such as sequence alignment, gene query, etc.

Data visualization could provide graphic representations for data and analysis results. Some underlying structures may be found through these visualized images. Thus, visualization is also an important analysis demand for researchers. The platform would provide the following visualization services.

- Frequency statistics visualization

  Users could check the visualized frequency statistics based on ages, gender, region, education levels, etc. They could also make filtering to the visualization according to specific attributes.

- Dimensionality reduction visualization

  Usually, a sample contains hundreds or thousands of attributes; thus, it is impossible to visualized the sample in a 2-D or 3-D graphic without dimensionality reduction. In order to achieve this type of visualization, the platform adopts linear and nonlinear dimensionality reduction methods to reduce the samples' dimension to 2 or 3, such that they could be visualized. Linear dimensionality reduction includes linear discriminant analysis (LDA), principal component analysis (PCA), etc. Nonlinear dimensionality reduction includes locally linear embedding (LLE), Laplacian eigenmaps, etc.

- Cohort study visualization

  Cohort study is a type of clinical study design. It is in the form of longitudinal study that aims to following a group of people for risk factor analysis. Therefore, the data changes with a time serial. The platform offers a dynamic cohort data visualization so that users could check the change of finite risk factors in the graphic format.

- Cross-sectional study visualization

  Cross-sectional study aims to provide data on the entire population under study. For this kind of data, the platform provides visualization option for researchers such that they could compare the data from different individuals by graphic format.

4.5 Implementation Technologies for Data Platform

In order to guarantee platform's safety and robust, some advanced technologies are adopted, which is illustrated in Fig. 4.3. The entrance module employs proxy system and load balance system, and the service module is implemented by Nginx, NodeJS, and Python. There are more complex issues to be concerned of when implementing data module and computing module; hence, the detailed technologies for these two modules are described below.

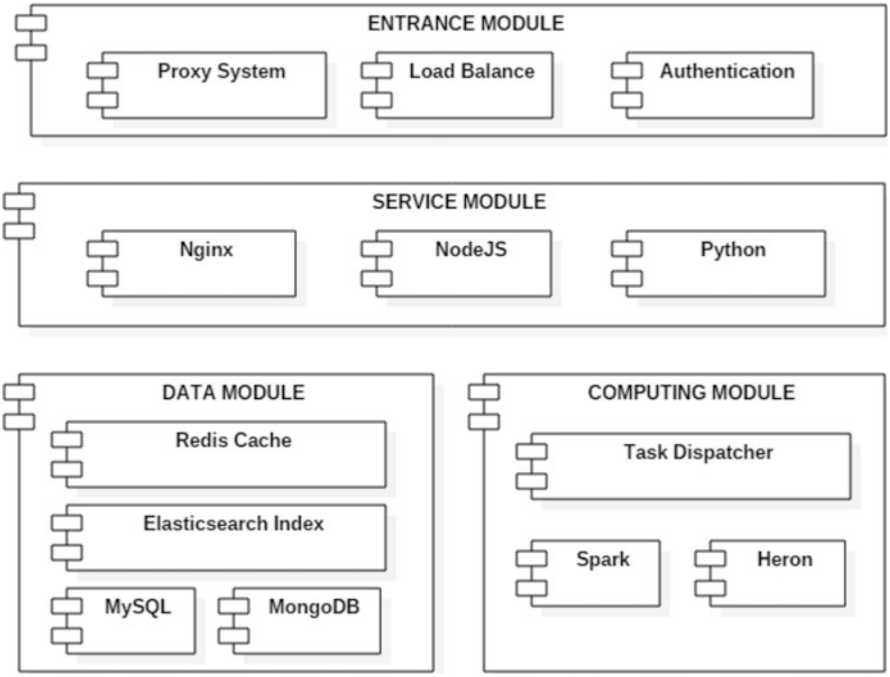

Fig. 4.3 Technologies adopted by the data platform. The entrance module employs proxy system and load balance system, and the service module is implemented by Nginx, NodeJS, and Python. In data module, MySQL and MongoDB are used to store data, while Redis Cache and Elasticsearch index are employed to improve efficiency. Task Dispatcher, Spark, and Heron are applied in computing module to improve the computation efficient

4.5.1 Storage

There are two issues needing to be concerned, including:

1. The platform supports various types of data, and these data could not be treated equally due to different formats. There are mainly three data formats, i.e., structural data, such as tables; unstructured data, such as unstructured text and image data; and semi-structured data, which contain both data types.
2. The platform needs support for high synchronizing access, and distributed storage should be adopted due to huge amount of data.

Only using purely relational database to store these data is unrealistic. Therefore, the framework in 1.3.2 illustrates that the platform adopts a specific system to solve these problems.

The storage consists of four parts:

1. Cache layer:

Redis is applied to cache users’ frequency data since it stores data in memory and supports numerous data types (including txt, number, row, dictionary, etc.).

This layer is used to accelerate user's data access and release the burden of storage system.

2. Index layer:

Elasticsearch is employed to provide index for txt files due to two reasons, one is that it can provide strong txt data index and composite index and the other is that it can accelerate the speed of index and release the burden of storage system.

3. Structural data storage:

The platform uses MySQL to store structural data. MySQL is the most well-known open-source relational database engine, which is stable, is low cost of training and establishment, and supports distributed database cluster with outstanding performance among open-source products.

4. Unstructured data storage:

The platform uses MongoDB to store unstructured data, including image data and document data. MongoDB supports master-slave distributed cluster storage and has strong querying function.

4.5.2 Computation

Computation module accomplishes all the computation task of the data platform. There are two issues that need to be concerned of for this module:

- Huge amount of computation tasks: computation technologies need to be adopted to improve the computing efficiency.
- Real-time demand: when utilizing the data analysis services, users may not wait for a long time for the results; this proposes real-time demand for computation module.

To tackle these two issues, the mixed computing framework is applied, including Spark and Heron. Spark is a streaming model-based platform, which is implemented by Java and provides multi-language interfaces. It is suitable for handling batch processing of big data and weak in real-time processing. Heron is a streaming model-based real-time platform, which could read and deal with huge amounts of data. All the temporal data analysis would be assigned to Heron. These two platforms are not independent. When requiring numeral amount of batch processing, Spark will be in charge.

Figure 4.4 illustrates the data flow in the computation module. Spark handles the whole data analysis for each day, which could be regarded as the preprocessing of data analysis. The newly coming data would be handled in Heron, and then the results would be combined with the results obtained by Spark.

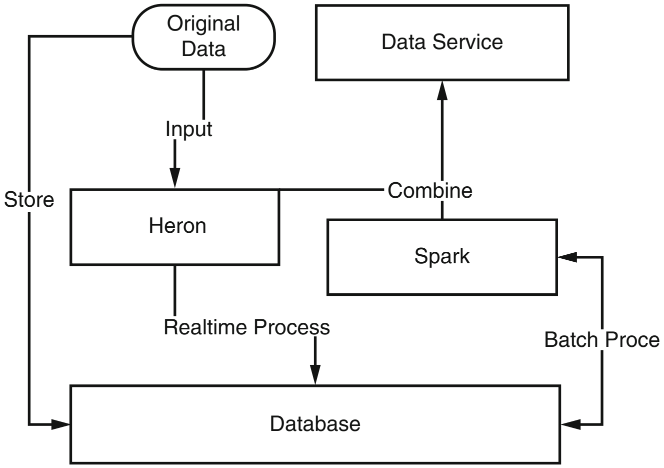

Fig. 4.4 Data flow in the computation module. Spark handles the whole data analysis for each day. The newly coming data would be handled in Heron, and then the results would be combined with the results obtained by Spark

4.6 Prevention of AD Based on Data Platform

The objective of establishing data platform is to facilitate AD's research and prevention. In this section, concrete strategies for advancing AD's prevention are introduced, including how to determine risk factors, conduct cognitive assessment, and improve cognition.

4.6.1 Risk Factor Determination

One of the important tasks for prevention is to identify AD's risk factors to guide prevention strategies. This task is based on the analysis of epidemiological investigation data. By calling the data analysis support service, the relationships between risk factors and the diseases could be determined.

Given a set of risk factors, the following analysis support could be applied for risk factor analysis. Student's t-test and χ^2 test are used to evaluate the risk factors' differences in midlife characteristics by dementia status; logistic regression analyses are used to validate the CAIDE risk score, which is calculated by the Kaplan-Meier method; C statistic is applied to evaluate the dichotomous outcome of ever/never dementia, where C statistic is the area under ROC curve; Cox proportional hazard models could be used for prediction modeling; partial least squares regression could be employed for risk factors selection in predicting AD [27].

4.6.2 Cognitive Assessment

Cognitive assessment is a technique used to assess the logical or reasoning abilities of human brains. It is critical in AD's prevention as it could be used for (1) screening for cognitive impairment, (2) differential diagnosis of cause, and (3) rating of severity of disorder or monitoring disease progression [28].

The assessments are mostly focused on attention, memory, visuospatial skills, and executive function abilities. The common methods for cognitive assessment are task-oriented assessment, which ask participants to complete a series of tasks like logical puzzles, matching numbers, etc. that necessitate the involvement of cognitive skills. Some classical tasks includes Prevention and Early Intervention Program for Psychoses (PEPP) cognitive assessment battery, Psychology Experiment Building Language (PEBL), Bhatia's Battery for Performance Test of Intelligence, neurocognitive test battery, cognitive drug research (CDR) computerized assessment system, etc. [29]

Some assessment tools have been developed based on these assessment tasks. The platform provides APIs to integrate these tools to facilitate AD's prevention. Some successful tools are described below:

- Alzheimer's Disease Pocketcard

  Alzheimer's Disease Pocketcard app can manage AD with confidence. It could help physicians and other healthcare professionals to take care of AD patients. Its highlights include (1) top ten signs of Alzheimer's disease; (2) latest information on detection, diagnosis, and management of Alzheimer's disease; (3) interactive tools to assess cognition and function, including the Mini-Cog, clock-drawing test, Saint Louis University Mental Status Exam (SLUMS), Functional Activities Questionnaire (FAQ), etc.; (4) annual wellness visit algorithm to help clinicians assess cognition more efficiently; (5) current diagnostic criteria, including the DSM-5 and the updated diagnostic criteria and guidelines for Alzheimer's disease from the National Institute on Aging and the Alzheimer's Association; and (6) education/support packets (PDF brochures) from the Alzheimer's Association that can be e-mailed directly to patients and caregivers.
- ANU Alzheimer's Disease Risk Index (ANU-ADRI)

  ANU-ADRI is an effective evidence-based tool to assess AD risk factors of people after 60 years old; it can provide personalized evaluation and exposure risk factors. It can help people know their current risk profile and areas, to reduce the risk of Alzheimer's disease in these areas. In addition, it can help the doctor to record their risk profile for the next consultation. ANU-ADRI can also be used for research projects to assess in reducing Alzheimer's risk [30].
- The CAIDE Risk Score (mobile application) App

  CAIDE dementia risk score app is a validated tool to predict the risk of dementia in his later years (20 years later) by involving age, education level, high blood pressure, high cholesterol, obesity, and lack of exercise. It can help users to reduce the modifiable risk factors and deferred cognitive impairment and dementia. Users could test their own personal risk and check the guidance for the modifiable risk. The app would recommend to consult a professional doctor if necessary. It also allows doctors to discuss preventive measures and monitor the reduction of risk [31].

4.6.3 Prevention Strategies and Cognitive Improvement

Cognitive assessment is to evaluate the cognitive status, while risk factor determination is to find the AD-related risk factors. Both strategies could provide guidance for AD prevention and improvement. The platform provides a section to maintain this guidance. Some typical guidance are introduced as follows.

Lifestyle can reduce the incidence of AD which contains physical exercise, cognitive ability, level of education, and social competence. The effect of aerobic exercise may come from the improvement of cardiovascular health and cardiovascular health. In the mouse models, environmental enrichment, including repeated exposure to novelty, has been shown to reduce amyloid burden associated with neuroanatomical and behavioral defects. Powerful social participation may help to reduce the likelihood of elderly peoples' suffering from dementia including AD. It may also help to slow down the development of symptoms [32].

These guidances may not be easy to access from the website; thus, the platform provides APIs to allow mobile phones to access these information. Some successful apps for cognitive improvement have been developed, e.g., ADcope. ADcope contains five modules: the first one is memory wallet, which consists of 30 images or sentences about the familiar family, places, and events; the second one is calendar, which is used to remind patients of daily activities; the third one is NFC tags, which are placed on drawers, doors, etc.; by clicking the tag, the app will display the contents of the drawer or the room; the fourth one is audio-assisted memory training module, which could train the patient's memory by repeating biographical information and asking questions about the information; and the last one is spaced retrieval exercise module, which is used to enhance patient's memory ability by two-phase exercise, assessing, and training [33].

The strategies for improving cognitive usually require users to accomplish serials of activities. How to monitor the activities is an issue needing to concern. Fortunately, many smart devices have been developed to monitor human bodies' physical or motion information. Some typical smart devices are described below:

1. Smart bracelet

   Smart bracelet uses wrist smart devices to measure the heart rate and obtain simplified heart rate zone continuously and automatically. Take the Fitbit smart bracelet as an example; it integrates training mode to record exercise, and people could view the details and summary of exercise data in the smart phone in real time. Besides Fitbit smart bracelet can record steps, distance, calories book, numbers of stairs, and the active time. It could also automatically monitor the sleep hours and sleep quality. Users could set a target, record diet, view progress, and analyze trends by mobile phone.

2. Smart clothes

   Smart clothes use flexible sensor technology to monitor the motion of the heart rate, breathing, and the main muscle groups of the body surface in real time. Take BodyPlus as an example; people can develop the most suitable fitness plan for their own according to their situation. BodyPlus equipped with powerful data analysis tools could sum up and analyze people's training and generate an analysis report for the heart and lung function and muscle stimulation situation.

3. Intelligent headset

   When users exercise with a headphone, the embedded motion sensor can monitor the body's temperature, perspiration, heart rate, and other indicators. Besides, earplugs' base frame is also equipped with acceleration sensors, making it achieve more precise motion data.

4. Intelligent insole

   Intelligent insole could be used to monitor human's respiration, heart rate, exercise, sleep, gait, and other information continuously.

5. Smart mattress

   Smart mattress could measure users' breathing and heart rate during sleep and send the collected data to smart phones wirelessly. The corresponding mobile application could show the depth of sleep, snoring time, etc., and assess the quality of people's sleep and give suggestions for improvement.

These smart devices need to communicate with smart phone for data processing and visualization. Some data process may be too complex to be run on the phones. The platform would provide APIs for these smart devices to upload their data for analysis. Through the data service module, the more comprehensive data analysis functions could be obtained.

4.7 Research of AD Based on Data Platform

The platform incorporates comprehensive data analysis methods to support AD's research. Some classical cases for these analysis methods' application are introduced below, including biomarkers extracted from medical images, diagnosis assistant, AD prediction, etc.

4.7.1 Biomarkers Extracted from Medical Images

Various types of medical images have been integrated in the platform, including magnetic resonance imaging (MRI), positron emission tomography (PET), etc. These images contain specific topologies, and biomarkers extracted from

AD/MCI's MRI image and PET image have been proven to be able to aid diagnosis [36].

Currently, medical images' biomarkers are mostly extracted manually, which requires professional background knowledge. Some doctors lacking in AD image knowledge may not diagnose AD accurately. These wrong diagnoses of AD would cause patients' unnecessary fear, despair, or even discrimination. Hereby, biomarker extraction becomes an important task for AD research and contributes significantly to the diagnosis and pathological study of AD.

The platform provides back-propagation (BP) neural network method for biomarkers derived from medical images. This method is one of the most famous machine learning modeling methods to recognize and classify features from medical image and has been widely used in the diagnosis systems of lung cancer. Users could utilize BP neural network method provided by the platform to extract biomarkers from MRI and PET. By establishing a multilayer feedforward neural network, the method uses numeral numbers of image data as training set and adjusts parameters based on mean square error to achieve biomarker extraction.

Data platform also supports convolutional neural network (CNN) for image analysis. CNN is a deep learning method, which adds a new step which imitates the way of human brain's processing signal for feature learning. CNN uses convolutional layer and dimension reduction layer to make the computer to extract biomarkers close to medical expert, which could improve the accuracy of extracting biomarkers from AD image data.

4.7.2 Diagnosis Assistant

Currently, AD diagnosis depends on the pathological diagnosis. An issue emerges that AD couldn't be definitely diagnosed before patients' death. Research shows that the misdiagnosis rate of AD range from 27 to 57%. Therefore, how to improve the accuracy of AD diagnosis is the most crucial problem of AD-related research.

Actually, diagnosis could be converted into machine learning classification problem, which requires two stages: training and classification. The former one is to analyze data set and establish classification models, and the latter one is to apply the new sample to the model and output the prediction value. In short, the whole process could be regarded as compressing patients' data into a number associated with AD so that the patients could be classified whether or not the patients suffer from AD [37].

The platform integrates various machine learning algorithms for AD diagnosis assistant, including support vector machine (SVM), relevance vector machine (RVM), back-propagation neural network, etc. The following introduce how to apply these algorithms on different types of data to make a diagnosis:

- *Diagnosis based on questionnaire data*

  For questionnaire data, random forest algorithm could be applied. Take Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) as an example; random forest is employed to generate the scale of ADAS-weighted tree. According to the ADAS-weighted tree, we can estimate whether the patient is normal, AD, or MCI. Compared with magnetic resonance imaging (MRI) or extraction of cerebrospinal fluid biomarker measurement, ADAS-weighted tree is considered to be more economical and useful. Moreover, the ADAS tree method won't cause great influence to the investigators' health.
- *Diagnosis based on image data*

  Support vector machine (SVM) is a widely used method for image-based diagnosis. By using biomarker extracted from MRI or PET image data, a classifier could be established based on SVM to classify the biomarker data and assign each sample a value as a label. These labels could be regarded as the diagnosis results.

  SVM uses kernel function to transfer the input space into a high-dimensional space, and then an optimal classified hyperplane is determined. Sometimes, dimension reduction is applied to reduce the computation complex before establishing SVM classifier model, which means some information may be discarded. This could introduce some errors. Hereby, the platform provides penalized logistic regression and coordinate-wise optimization methods to overcome the disadvantage of dimension reduction.

  Furthermore, relevance vector machine (RVM) algorithm is provided for AD classification, which is based on sparse Bayesian framework. RVM first transfer the input data into a high-dimensional space by choosing a proper kernel function such as Gaussian kernel function, Laplace kernel function, polynomial kernel function, etc. Among which, Gaussian kernel is the most widely used kernel. Then RVM needs to initialize the hyper-parameter and obtain the ideal model of the classifier by iteration. Compared to SVM algorithm that is based on structural risk minimization, RVM may have a stronger generalization ability.
- *Diagnosis based on multimode data*

  Previous methods are based on single type of data. Sometimes, multiple types of data are available. Combining various types of data to make a diagnosis could achieve more accuracy results [38], for example, any two types or all types of questionnaire data (MMSE, ADAS-Cog, CDR), image data (MRI, PET), gene data, and biomarker data (CSF) could be combined. Therefore, the platform provides options for multimode data analysis.

  One method is to integrate both MRI image data and MMSE scale to make a diagnosis. By building resting-state brain functional networks, we can extract the abnormal network nodes' properties and train the SVM classifier based on voxel-based morphometry (VBM).
- Another method is to combine image data (FDG-PET image, MR scan image), extracted biomarker data of cerebral spinal fluid (CSF), and questionnaire data (MMSE, ADAS-Cog) to make a classification of AD, MCI, and NC. This is also a SVM-based method. At first, the algorithm processes feature extraction and

feature selection from image data and calculates the kernel matrix based on selected feature. For CSF data, the algorithm directly calculates the kernel matrix without feature selection. Then the method combines all kernel matrixes (MRI, PET, CSF) to process classification by using SVM algorithm. Comprehensive studies show that this method could achieve higher accuracy than the other methods based on single type of data.

4.7.3 AD Prediction

Another important AD research is to make prediction about the progress of the disease, e.g., the onset time, disease exacerbation, transformation of disease symptoms, etc. This relies regression analysis of cohort clinical data. Thanks to the development of epidemiological investigation, many cohort data have been collected, which make the AD prediction possible [39].

The prediction could be done on single type of data, but this may not achieve inspired performance. Therefore, the platform supports a multimode prediction method based on combined different kinds of data. This method not only makes regression analysis from collected clinical data but also combines with the mentioned classification method in the previous chapter. Besides, we use the corresponding multimodal support vector regression (SVR) algorithm as the basis. This method not only predicts clinical data through the regression method but also predicts the 2-year changes of MMSE and ADAS-Cog scores to help validate the prediction results.

4.8 Conclusion

Data share is an important strategy to advance AD area. Therefore, data platform is designed to integrate data from various institutes to enable big data analysis, which could accelerate AD's research and prevention. Besides, comprehensive services are provided to simplify searchers' analysis task for data. Prevention and research cases based on data platform illustrate the importance of the data platform. Those who devote to AD area should turn to the data platform for data share and analysis support.

Acknowledgments This study was supported partially by Anhui Provincial Key Technologies R&D Program under Grant No. 1704e1002221, the National Natural Science Foundation of China (NSFC) under Grant No. 71661167004 and the Programme of Introducing Talents of Discipline to Universities ("111 Program") under Grant No. B14025.

References

1. Hebert LE, Weuve J, Scherr PA, Evans DA (2013) Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. Neurology 80(19):1778–1783
2. Prince M, Wimo A, Guerchet M, Ali GC, Wu YT, Prina M. (2015) The global impact of dementia: an analysis of prevalence, incidence, cost and trends. World Alzheimer Report
3. Wimo A, Jönsson L, Bond J, Prince M, Winblad B, International AD (2013) The worldwide economic impact of dementia 2010. Alzheimers Dement 9(1):1–1
4. Seshadri S, Beiser A, Kelly-Hayes M, Kase CS, Au R, Kannel WB, Wolf PA (2006) The lifetime risk of stroke estimates from the Framingham study. Stroke 37(2):345–350
5. Mahmood SS, Levy D, Vasan RS, Wang TJ (2014) The Framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. Lancet 383(9921):999–1008
6. Alzheimer's A (2015) 2015 Alzheimer's disease facts and figures. Alzheimers Dement 11 (3):332
7. Shimmer, Wireless Sensing Solutions for werable applications, online at: http://shimmer-research.com/
8. Souillard-Mandar W, Davis R, Rudin C, Au R, Libon DJ, Swenson R, Price CC, Lamar M, Penney DL (2016) Learning classification models of cognitive conditions from subtle behaviors in the digital Clock Drawing Test. Mach Learn 102(3):393–441
9. Zlokovic BV (2011) Neurovascular pathways to neurodegeneration in Alzheimer's disease and other disorders. Nat Rev Neurosci 12(12):723–738
10. Honig LS, Kukull W, Mayeux R (2005) Atherosclerosis and AD analysis of data from the US National Alzheimer's coordinating center[J]. Neurology 64(3):494–500
11. Shaw LM, Vanderstichele H, Knapik-Czajka M et al (2009) Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects[J]. Ann Neurol 65 (4):403–413
12. Jack CR, Bernstein MA, Fox NC et al (2008) The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods[J]. J Magn Reson Imaging 27(4):685–691
13. Hodes RJ, Buckholtz N (2016) Accelerating medicines partnership: Alzheimer's disease (AMP-AD) knowledge portal aids Alzheimer's drug discovery through open data sharing [J]. Expert Opin Ther Targets
14. Simmons A, Westman E, Muehlboeck S et al (2011) The AddNeuroMed framework for multi-centre MRI assessment of Alzheimer's disease: experience from the first 24 months[J]. Int J Geriatr Psychiatry 26(1):75–82
15. Lovestone S, Francis P, Kloszewska I et al (2009) AddNeuroMed—the European collaboration for the discovery of novel biomarkers for Alzheimer's disease[J]. Ann N Y Acad Sci 1180 (1):36–46
16. Partch AB, Laufer D, Valladares O et al (2015) Nia genetics of Alzheimer's disease data storage site (NIAGADS): 2015 update[J]. Alzheimers Dement 11(7):P362
17. Pérez-Cano R, Vranckx JJ, Lasso JM et al (2012) Prospective trial of adipose-derived regenerative cell (ADRC)-enriched fat grafting for partial mastectomy defects: the RESTORE-2 trial[J]. Eur J Surg Oncol (EJSO) 38(5):382–389
18. Thal LJ (2004) The Alzheimer's disease cooperative study in 2004.[J]. Alzheimer Dis Assoc Disord 18(4):183–185
19. Zhao Y, Hu Y, Smith JP et al (2014) Cohort profile: the China health and retirement longitudinal study (CHARLS).[J]. Int J Epidemiol 43(1):61
20. Cerami E, Gao J, Dogrusoz U et al (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data[J]. Cancer Discov 2(5):401–404
21. CommonMind Consortium Launched as a Public-Private Effort to Generate and Broadly Share Molecular Data on Neuropsychiatric Disease[J] (2012) Biomedical Market Newsletter
22. Day K, Mcguire L, Anderson L (2009) The CDC healthy brain initiative: public health and cognitive impairment[J]. Generations 33(1):11–17

23. States A (2013) The healthy brain initiative: the public health road map for state and national partnerships, 2013–2018[J]. Aging/physiology/united States
24. Jiang J (2012) Discussion on the information Organization of the Scientific Data Sharing Platform-Take the National Scientific Data Sharing Platform for population and health as an example[J]. J Inf Resour Manag
25. Centers for Disease Control and Prevention (2014) Health indicators warehouse[J]. Obesity in children and adolescents aged. 2–19
26. Parkinson J (2008) Common neurological disorders: Parkinson’s disease[J]. Br J Healthc Assistants January
27. Exalto LG, Quesenberry CP, Barnes D, Kivipelto M, Biessels GJ, Whitmer RA (2014) Midlife risk score for the prediction of dementia four decades later. Alzheimer’s Dement: J Alzheimer’s Assoc 10:562–570. doi:10.1016/j.jalz.2013.05.1772
28. Woodford H, George J (2007) Cognitive assessment in the elderly: a review of clinical methods. QJM 100:469–484
29. Singh M, Sachdeva S (2014) Cognitive assessment techniques. Int J Inf Technol Knowl Manag 7(2):108–118
30. Anstey KJ et al (2014) A self-report risk index to predict occurrence of dementia in three independent cohorts of older adults. The ANU-ADRI. PLoS One 9:e86141
31. Sindi S, Calov E, Fokkens J et al (2015) The CAIDE dementia risk score app: the development of an evidence-based mobile application to predict the risk of dementia. Alzheimers Dement 1:328–333
32. Selkoe DJ (2012) Preventing Alzheimer’s disease. Science 337(6101):1488–1492
33. Zmily A, Mowafi Y, Mashal E (2014) Study of the usability of spaced retrieval exercise using mobile devices for Alzheimer’s disease rehabilitation. JMIR mHealth uHealth 2(3):e31
34. Bloom GS (2014) Amyloid-β and tau: the trigger and bullet in Alzheimer disease pathogenesis. [J]. JAMA Neurol 71(4):505–508
35. Pastor P, Roe CM, Villegas A et al (2003) Apolipoprotein Eε4 modifies Alzheimer’s disease onset in an E280A PS1 kindred[J]. Ann Neurol 54(2):163–169
36. Gudbjartsson H, Patz S (1995) The Rician distribution of noisy MRI data[J]. Magn Reson Med 34(6):910
37. Escudero J, Ifeachor E, Zajicek JP et al (2013) Machine learning-based method for personalized and cost-effective detection of Alzheimer’s disease[J]. IEEE Trans Biomed Eng 60 (1):164–168
38. Zhang D, Wang Y, Zhou L et al (2011) Multimodal classification of Alzheimer’s disease and mild cognitive impairment[J]. NeuroImage 55(3):856–867
39. Hinrichs C, Singh V, Xu G et al (2011) Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population[J]. NeuroImage 55 (2):574–589