HHS Public Access

Author manuscript

Int J Med Inform. Author manuscript; available in PMC 2018 October 01.

Published in final edited form as:

Int J Med Inform. 2017 October ; 106: 48–56. doi:10.1016/j.ijmedinf.2017.07.002.

Advancing Alzheimer's Research: A Review of Big Data Promises

Rui Zhang, PhD[*],

Institute for Health Informatics and College of Pharmacy, University of Minnesota, Minneapolis, Minnesota

Gyorgy Simon, PhD, and

Institute for Health Informatics and Department of Medicine, University of Minnesota, Minneapolis, Minnesota

Fang Yu, PhD, RN, GNP-BC, FGSA, FAAN

School of Nursing, University of Minnesota, Minneapolis, Minnesota

Abstract

Objective—To review the current state of science using big data to advance AD research and practice. In particular, we analyzed the types of research foci addressed and corresponding methods employed and study findings reported using big data in AD.

Method—Systematic review was conducted for articles published in PubMed from January 1, 2010 through December 31, 2015. Keywords with AD and big data analytics were used for literature retrieval. Articles were reviewed and included if they met the eligibility criteria.

Results—Thirty-eight articles were included in this review. They can be categorized into six research foci: diagnosing AD or mild cognitive impairment (MCI) (n=10), predicting MCI to AD conversion (n=13), stratifying risks for AD (n=5), mining the literature for knowledge discovery (n=4), predicting AD progression (n=2), describing clinical care for persons with AD (n=3) and understanding the relationship between cognition and AD (n=3). The most commonly used datasets are Alzheimer's Disease Neuroimaging Initiative (ADNI) (n= 16), electronic health records (EHR) (n=11), MEDLINE (n=3), and other research datasets (n=8). Logistic regression (n=9) and support vector machine (n=8) are the most used methods for data analysis.

Conclusion—Big data are increasingly used to address AD related research questions. While existing research datasets are frequently used, other datasets such as EHR data provide a unique, yet under-tapped opportunity for advancing AD research.

[*]Corresponding author. Address: 330 Diehl Hall, 505 Essex Street SE, Minneapolis, MN 55455.

Conflict of interest

The authors declare that they have no competing interests.

Authors' contributions

RZ, GS and FY were responsible for the study design. RZ performed the review selection, and FY and GS assessed the quality of review. RZ, GS and FY analyzed the literature and drafted this manuscript. All authors read and approved the manuscript.

Keywords

healthcare big data; healthcare data analytics; Alzheimer's disease; Alzheimer's Disease Neuroimaging Initiative; electronic health records

1. Introduction

With the silver tsunami (that is, an aging workforce) sweeping the globe, Alzheimer's disease (AD) is becoming an endemic due to its disproportional afflictions on older adults who are 65 years old or older. AD is the most common type of dementia constituting 60–80% of all dementias. As of 2013, an estimated 44.4 million people had dementia worldwide, and this number is projected to reach 75.6 million in 2030 and 135.5 million in 2050 with most of the increase occurring in developing countries (1). In the U.S., 5.3 million Americans had AD in 2015. Of those, 5.1 million are older adults, which will be almost tripled to 13.8 million by 2050 (2). The exponential growth in AD prevalence will not lessen unless medical breakthroughs to prevent or cure AD are developed in the next few decades. As the sixth leading cause of death in the U.S., AD is the only one that cannot be prevented, slowed, or cured. While deaths from other causes have decreased substantially in the past few decades, deaths from AD have increased significantly (2). Dementia is also one of the most expensive chronic diseases to the society with $604 billion expenses worldwide in 2010 (1). AD alone is estimated to cost the American society $226 million in 2015. Moreover, AD affects the whole families and social networks. In 2014, families and friends provided people with AD 17.9 billion hours of unpaid care which was valued at $217.7 billion in the U.S. Caregiving exerts heavy physical and emotional tolls on those caregivers by causing new diseases or exacerbation of existing conditions which amounted to another $9.7 billion in health care (2). Hence, it's paramount to develop effective interventions to prevent AD from occurrence, slow down AD progression once it occurs, and improve quality of life and care for people and families who are affected by AD (3).

In particular, we reviewed the current state of science to generate broad themes of research foci addressed using big data in the past 5 years inductively, analyzed the corresponding analytical methods employed, and synthesized the study findings. Big data refer to "large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information" (4). Recently, the rapid adoption of healthcare information technology has dramatically accumulated vast amounts of heterogeneous healthcare big data. Big data research has substantially influenced many fields in biomedical and healthcare domains, such as cancer (5,6), diabetes (7,8) and heart failure (9). However, it's unclear how healthcare big data have been used in AD research. In this paper, we address three questions:

1. What research foci have been addressed using big data in AD research?

   For each research focus, we further evaluated:

2. Which databases or datasets were used?
3. What were the primary research methods and key findings?

2. Background

The term “big data” has been frequently used in many fields, and its definition is always evolving. The original characteristics of the “big data” are defined as the three Vs: volume, variety, and velocity (10). The first feature is that the volume of data is aggregating dramatically in the past decade. For example, US healthcare system has already reached 150 exabyte (10^{18}) in 2013 (11). Big data in healthcare will soon reach the zettabyte (10^{21}) and later even yottabyte (10^{24}). The second characteristic, variety, refers the heterogeneity nature of big data. Data can be collected from many sources, including microarray data, imaging data, structured data (e.g., medication, diagnosis), and unstructured data (e.g., clinical notes). The third characteristic, velocity, indicates the speed of generating data. For example, the current sequencing techniques can produce billions of sequence data daily. Electronic health record (EHR) systems can generate millions to billions of medical records per day. Besides these characteristics, other three features were also considered: variability, veracity, and value (12). Variability refers to the consistency of data over time. Veracity is vital for big data since data are sometimes from uncontrolled environments, such as less reliable ambulatory measurements. Value of the big data for healthcare and patient can be obtained when the challenges of big data analytics can be addressed. Despite this new definition for big data, there are no agreed upon definition and properties for big data in health care research. In this review, we defined big data as complex and heterogeneous in magnitude, which are difficult to collect and manage in traditional ways, including: 1) datasets were collected by more than one site, such as the AD Neuroimaging Initiatives (ADNI) to aggregate data; 2) patient data from one or more EHR systems(13) ; 3) integration of heterogeneous data sources, such as clinical measurements, imaging data, or behavior data; or 4) biomedical literature (i.e., MEDLINE citations) from which new information or knowledge can be discovered for AD research.

AD is a progressive, irreversible, neurodegenerative condition that attacks the brain structure (e.g., neurons and their milieu) and results in brain functional loss. AD cerebral neuropathology includes β-amyloid accumulations outside neurons disrupting neuron’s environments and tau hyperphosphorylation that destroys neurons from within. Clinical diagnosis of AD is typically made when there are two or more domains of cognitive impairment which isn’t attributable to other causes (e.g., infection) and is severe enough to cause functional decline in occupational, social, instrumental, and basic activities of daily living. AD-resulted cognitive impairment typically manifests as memory loss and impaired executive function, visuospatial function, or language. Cognitive impairment, functional decline, and behavioral and psychological symptoms of dementia (BPSD) are considered the triad symptoms of AD. Definitive diagnosis of AD can only be made upon death via brain biopsy (14). Although the accuracy of clinical diagnosis of AD is comparable to other conditions at 80–90%, only 45% of people with AD or their caregivers were ever told of the AD diagnosis (2).

2.1 Research Progress in AD

The past two-to-three decades have witnessed tremendous research efforts that have led to our increased understanding of AD, which can be summarized into six broad themes;

however, it's important to recognize that research effort and funding for each theme vary substantially: 1) Identifying risk factors (e.g., epolipoprotein, cardiovascular risk factors) and testing interventions (e.g., physical activity) to diagnose AD from occurrence; 2) Predicting MCI to AD conversion using imaging, neuropsychological data; 3) Stratifying AD risks with associated factors; 4) discovering novel biomarkers or potential AD drugs from diverse resources (e.g., literature) ; 5) tracking AD progression using imaging, cerebrospinal fluid, and blood biomarkers (e.g., Pittsburg compound B); 6) addressing caregivers and caregiving issues (e.g., caregiver burden, dealing with behavioral and psychological symptoms of dementia); and 7) understanding the relationship between cognition and AD (15–18).

The concerted scientific insights and technological advances have borne fruit of three guidelines in 2011 to improve AD diagnosis (14) and identify prodromal stages of AD as mild cognitive impairment (MCI) due to Alzheimer's (19), and preclinical (presymptomatic) Alzheimer's (20) to prevent AD in 2011 (14,19,20). While clinical trials to find a cure for AD have failed miserably at an unsurmounted failing rate of 99.6% from 2002 to 2012 (3) (3), substantial progress has been made in non-pharamcological treatments for improving symptoms and quality of life in people with dementia and their caregivers. However, those findings have not been integrated into traditional health care practice and their translatability or applicability are yet unknown.

2.2 Sources of big AD data

The importance of big data to enhance the AD research has been recognized by the AD research community. In 2014, the Global CEO Initiative on AD (CEOi), in collaboration with Sage Bionetworks and IBM's DREAM project launched AD Big Data Challenge at the White House to advance the global effort for diagnosis techniques and identify new AD biomarkers through open source data (21). The open source data from persons with AD were provided by the North American Alzheimer's Disease Neuroimaging Initiative (ADNI), and the European's AddNeuroMed Study. The ADNI dataset is longitudinal multicenter study to develop clinical genetic and biomedical biomarkers for AD early detection (22). The study began in 2004 and experienced three phases ADNI1, ADNI GO, and ADNI2. Right now, ADNI study has enrolled over 1600 subjects from cognitive normal to AD. In addition, the AD research community began to recognize the potential of the EHR systems which have been increasingly adopted worldwide with exponential aggregation of patient data in the recent years. EHR is also considered a good source of big data and can provide rich real world information for AD research with the appropriate computational methods (13). Mayo Clinic Study of Aging (MCSA) was designed in 2008 for a prospective population-based study of normal cognitive aging, MCI and dementia (63). Through random sampling and criteria evaluation, 2,719 subjects were identified through a review of their medical records from a population of Olmsted County in Minnesota of the United State. Among these, through personally interview for 2,050 participants and telephone interview for 669 participants, 402 subject with dementia were identified. Their clinical characteristics, including coronary heart disease, diabetes, hypertension, *etc*, were evaluated during the interview. ZARAgoza DEMentia DEPression (ZARADEMP) study is cohort study for about 5,000 individuals enrolled in Zaragoza of Spain. HPC, beside containing multiple image

modalities, also contains behavioral and genetic information. Table 1 listed the facts related to some selected databases used for AD research in the world.

However, it's unclear how big data have been used in AD research and what the state of the science is currently. Hence, the purpose of this paper was to review the current state of science using healthcare data analytics to advance AD research. In particular, we analyzed the types of research foci addressed and corresponding methods employed and study findings reported using big data in AD.

3. Methods

We conducted a literature search in PubMed, the most comprehensive reference database in healthcare, from January 1, 2010 through December 31, 2015. We used the search term "Alzheimer Disease" or "Alzheimer's Disease" in title or abstract, and combined each of those two terms with another search term listed under in the Data or Data Analysis category in Table 2. We retrieved 141 articles based on the search. The title and abstract of each article was reviewed to determine if the study met the inclusion criteria: 1) publication in English; 2) available full-text; and 3) human subjects. Next, articles meeting the following exclusion criteria were excluded: 1) not related to AD; 2) lacking a clinical emphasis (e.g., AD diagnosis, treatment, or management); 3) genome-wide association study or analysis on microarray data; and 4) editorial, commentary, review or conference summary. At a result, 38 articles met the eligibility criteria and were included in this review.

4. Results

This section summaries synthesized findings of the reviewed articles on data analytics based on the AD research question categories.

Analysis of the 38 articles showed that six main research foci have been addressed using big data: diagnosing AD or MCI (n=10), predicting MCI to AD conversion (n=13), stratifying risks for AD (n=5), mining the literature for knowledge discovery (n=4), predicting AD progression (n=2), describing clinical care for persons with AD (n=3), and understanding the relationship between cognition and AD (n=3). The most commonly used datasets are ADNI (n= 16), followed by EHR (n=11), MEDLINE (n=3), and other research datasets (n=8). Logistic regression (n=9) and support vector machine (n=8) are the most used methods for data analysis. The dataset, analytics, main findings, and future research directions for each of the six main research foci are synthesized below.

4.1 Diagnosing AD or MCI

Ten out of the 38 articles focused on early detection of AD and MCI using cerebrospinal fluid (CSF) biomarkers, cognitive and memory measurements (e.g., Mini Mental State Examination (MMSE) scores), and imaging results (e.g., magnetic resonance imaging (MRI), positron emission tomography (PET)) (Table 3). The most frequently used dataset is the ADNI database (n=7). The 10 articles tested different analytical methods for analyzing big data to construct biomarker sets that have diagnostic value for AD and/or MCI by using serum, CSF, imaging, or physical biomarkers. Five articles used SVM with various feature

sets to classify AD, MCI and HC. Two applied logistic regression, and others used random forest classifier or statistic methods. They discovered novel biomarkers, imaging features, anatomical features, other phenotypes (e.g., semantic fluency and eye movements) associated with classification.

Van Gils *et al.* first identified biomarker subsets that could provide a reliable and early detection of AD prior to any major clinical signs (23). Li *et al.* found anatomic features to discriminate AD or MCI with HC (24). Yang *et al.* used volumetric and shape features from MRI scans to diagnose AD and MCI patients (25). Mangialasche *et al.* applied a multivariate data analysis technique to differentiate AD and MCI subjects from HC subjects (26). Kohannim *et al.* combined brain imaging and other biomarkers to classify ADNI subjects as AD, MCI and NC (27). Other biomarkers to detect AD and MCI included semantic fluency and eye movement. Clark *et al.* constructed random forest classifiers using latent information in semantic fluency word lists to predict cognitive and functional decline (28). Lagun *et al.* detected MCI using eye movement characteristics such as fixations, saccades, and refixations during the visual paired comparison (VPC) task (29). Casanova *et al.* introduced AD pattern similarity (AD-PS) scores, estimated by structural MRI and cognitive test data in ADNI to conduct classification (30). A multi-model multi-task learning (M3T) method was used to classify patients with value of AD, MCI or HC (32). A diagnostic clinical decision support system (CDSS) for early diagnosis of AD was implemented (33). The method performed slightly worse than benchmark method when it was applied to publically available medical datasets.

4.2 Predicting MCI to AD conversion

Thirteen studies developed various methods to predict the conversion from MCI to AD using MRI (n=4), electroencephalography (n=1), clinical data (n=4), and a combination of those (n=4). The datasets used included ADNI (n= 7), Zaragoza Dementia and Depression Project (ZARADEMP), a longitudinal epidemiologic study in Spain (34) (n=1), and AddNeuroMed study data (n=1). Other studies collected imaging data, neuropsychological and clinical data from enrolled patients (n=5) (Table 4). Most used method is regression (n=6), followed by SVM (n=2), Bayesian network and deep learning. Various features including imaging data, EEG biomarkers, neuropsychological evaluation tests were found to be associated with MCI to AD conversion.

Five studies developed different algorithms to predict MCI to AD conversion using MRI imaging data in ADNI. Two studies demonstrated that Bayesian network accurately (0.75) differentiated MCI converters from non-converters (35,36). Liu *et al.* improved the predication of MCI to AD conversion using local linear embedding (LLE) (37). Hinrichs *et al.* designed a multi-kernel learning (MKL) framework (38) and later applied Bayesian Gaussian process logistic regression (GP-LR) models to differentiate MCI patients from HC and AD patients (39). Deep learning techniques were applied to classify various stages of AD progression using MRI scans from ADNI database (40). Mattila *et al.* designed a statistical model, the Disease State Index (DSI), which could accurately predict conversion from MCI to AD (33).

In addition to ADNI data, Costafreda *et al.* predicted conversion from MCI to AD based on hippocampal morphology in AddNeuroMed (41), a longitudinal multi-site study of biomarkers for AD in the United Kingdom (42). In addition to imaging biomarkers, one study found that the six electroencephalography biomarkers on electroencephalography could predicte MCI to AD conversion in the Alzheimer's Center in Netherlands (43). Two studies used clinical tests to predict MCI to AD conversion. Pozueta *et al.* found that combining MMSE and California Verbal Learning Test Long Delayed Total Recall can predicate MCI to AD conversion (44). A multi-model multi-task learning (M3T) method was also used to predict clinical variables including MMSE and AD Assessment Scale-Cognitive Subscale (ADAS-Cog) (32).

Another four studies evaluated the predictive values of combined biomarkers and clinical data for MCI to AD conversion. Gomar *et al.* found that an episodic memory measure (i.e., AVLT Trial 5) and Clock Drawing test were the best predictors for MCI to AD conversion (46). Alegret *et al.* reported that semantic fluency tests and neuropsychological test results were significantly associated with the speed of conversion from MCI to AD (47). Runtti *et al.* created a disease stage index (DSI) value to classify MCI to AD converters (48).

4.3 Stratifying risks for AD

Five studies stratified AD risks using different methods. The datasets used included ZARADEMP (n= 1), National Alzheimer's Coordinating Center database (n=1), EHR data (n=1) Wisconsin Registry for Alzheimer's Prevention (WRAP), a prospective longitudinal study which began in 2001 (49) (n=1) and Ginkgo Evaluation of Memory data (n=1) (Table 5). Cox proportional hazard model were commonly used.

Gracia-Garcia *et al.* used multivariate regression method to analyze data from ZARADEMP and reported that severe depression significantly increases the risk of AD (36). Li *et al.* identified a significant association between erythrocyte sedimentation rate and AD using EHR data and VARiant Informing MEDicine (VARIMED) (50). Chang *et al.* created a new measure based on stochastic gradient descent to predict potential AD onset based on familial AD patterns (51). Rosenberg *et al.* found that neuropsychiatric symptoms (e.g., depression and anxiety) in MCI were associated with increased risk of dementia and AD (52). Last, Yasar *et al.* found that diuretic, angiotensin-1 receptor blockers, and angiotensin-converting enzyme inhibitors use was associated with reduced risk for AD (53).

4.4 Mining the literature and resources for knowledge discovery

Four studies examined the published literature for knowledge discovery in AD. The datasets used included the MEDLINE (n=2), a combination of review papers, MEDLINE, reports and databases (e.g., AD & Frontotemportal Dementia Mutation database, Gene Ontology database, NCBI Gene Expression Omnibus, Disease Database) (n=1), and a combination of MEDLINE and protein interaction data in the Online Predicted Human Interaction Database (OPHID) (n=1) (Table 6). Most of these studies used text mining or natural language processing techniques. Most studies focus on generate novel knowledge such as potential biomarkers or candidate AD drugs.

One study mined MEDLINE to generate 500 hypotheses (e.g., Tau and Amyloid-beta as potential biomarker candidates in relation to AD), which were then evaluated by the AlzSWAN, a comprehensive database containing expert curated AD-related hypotheses (54). The second study discovered 25 candidate AD biomarkers from diverse resources, including MEDLINE, AD research forums and related gene and disease databases (55). The third study generated AD-related proteins connectivity maps using protein interaction database and MEDLINE (56). The fourth study described a system which can identify highly relevant (84.5% accuracy) AD-related sentences from MEDLINE (57).

4.5 Predicting AD progression

Two studies evaluated the progression of AD. The dataset used included clinical records (n=1) and ADNI (n=1) (Table 7). The first study found that poor performance on the Trail making Test-A significantly predicted faster cognitive decline (58). The second study predicted cognition using clinical measurements from Mayo clinic datasets (59).

4.6 Describing clinical care for subjects with AD

Three studies evaluated the clinical care of persons with AD using EHR data. However, each study focused on different aspects of clinical care using different methods (Table 8). The first study showed that the AD diagnosis was associated with significant increases in primary and secondary care resources utilization (60). The second study showed that visits to dementia or mental health clinics increase the odds of receiving anti-dementia, antidepressant, and antipsychotic medications (61). The last study evaluated AD as an independent risk factor for hip fractures (62).

4.7 Understanding the relationship between cognition and AD

Three studies investigated the relationship between cognition and AD (Table 9). Rich data sources such as ADNI and MCSA have given rise to analyses that combine multiple data types. While image analysis is outside the scope for this survey, in this section, we highlight some examples of integrated analyses of multiple modalities including image and other data types. The MCSA data set combines image data with clinical (EHR) data allowing for detailed analyses of vascular disease and AD pathology (68). A study on an AD cohort with 20 years of follow-up suggests that education acts as a buffer against the clinical decline in AD (70). Such buffer, that allows an individual to maintain cognitive function despite AD pathology is called cognitive reserve (71). Combination of image data with behavioral data and genetic information in ADNI and in MCSA allows for assessing the change in cognitive reserve in the presence of genetic and other biomarkers (69).

5. Discussion

Given the lack of effective treatments for persons with AD and only 45% of people with AD were diagnosed (2) much emphasis has been placed on timely diagnosis of AD and early identification of people who are at heightened risk for AD such as those with MCI. As a result, 9 articles focused on diagnosing AD or MCI and 15 articles were on predicting MCI to AD conversion using biomarkers and/or clinical data. Each analytical method seemed to be able to diagnose AD or MCI or predict MCI to AD conversion with a high accuracy.

Besides the obvious clinical significance, the focus on diagnosis can also be explained from a technical perspective: diagnostic accuracy and prognosis (including progression from MCI to AD) are key areas that are commonly recognized as the strength of data science and Big Data analytics in general (64).

The prerequisite for Big Data analytics, and essentially any data-intensive research for that matter, is data availability. As a result, immense effort has been devoted to the creation of research datasets, such as ADNI, AddNeuroMed study, MCSA, and Ginkgo Evaluation of Memory study, but these data sets still contain relatively small patient cohorts when compared with the size of big data in general. Big Data analytics is built upon the premise that small differences, that can only be modeled by complex techniques, can develop into large differences over long follow-up periods. When we have longitudinal data or when we combine multiple facets of the same cohort (e.g. image data with EHR data), these small differences are more likely to be captured in the combined data. The above data sources may be small in the number of patients, but they have volume (follow-up), variety (different data modality), can support Big Data-type multi-modal analyses and thus they can be considered Big Data. The majority of the included studies extracted their datasets from these existing research databases, but a few studies used large cohorts such as the ZARADEMP cohort or EHR data.

While the excitement for Big Data clinical research is still rising, somber voices pointing out the pitfalls of Big Data are appearing (65). With Big Data methods modeling small differences, the question of validity is of paramount importance. Rarely has any big data research (AD or general medicine) result been tested in another population for replicating and validating the findings. This issue is particularly relevant in the case AD research, where studies rely on a wide variety of occasionally disparate data sources (e.g. socio-economic or education-related data), which may not be available to the general AD research community.

A third issue concerns translation of the research findings into practice. The nature of a research database is different from that of real-world data, such as EHR data in its completeness and cohort representation. For example, the research database, such as ADNI, has complete data in all research oriented variables, while EHRs are mainly used for documenting clinical care for persons with AD and not for research. The fragmentation of patient care is reflected in the incompleteness of EHR data (66) which stands in sharp contrast with the completeness of data in research databases. In addition, cohort representation in research databases is often not reflective of the realworld patients due to the restricted eligibility rules on recruiting patients (31, 67). While the current developed models in research data may miss confounding factors which can only detected in unselected patients in research databases. Thus, using a well-established research dataset such as the ADNI provides an ideal milieu for addressing a targeted research focus, it’s unclear if and to what extent the results created under an “ideal” situation will hold out for real life patient data such as those captured in EHR for routine clinical visits. Replicating those findings in real life could be especially challenging if certain data are not clinically collected (e.g., CSF biomarkers) and collecting those data places individuals at high risk for adverse events (e.g., infection) or at a cost that might not be clinically justifiable. These issues may hinder the translation and application of the computational methods and research findings into practice.

Although Big Data has long passed its infancy in the field of general data science, the current state of AD resembles the early stages of Big Data research. The key prerequisite of large, multi-site data repositories such as ADNI has been created. Many studies have generated promising findings on either the performance of the computation methods or the novel biomarkers and confounders. However, this promise is unfulfilled yet. Several factors, including external validation of the findings and questions surrounding the applicability of findings from "ideal" research data to real-world data hinder the translation of these findings into practice. While many data elements (such as imaging data for health patients, genetic data, and detailed socioeconomic data) will likely remain in the research domain for the near future, exploring the secondly use of EHR data to overcome many of these hindrances appears the logical next step.

The implications of our findings include: 1) address each of the six identified research foci with more studies and especially using EHR data; 2) replicate findings from each foci in different data resources, feature selections, and analytics; 3) validate the findings from studies using research datasets with EHR data; and 4) focus on the clinical outcomes. Findings from our review suggest that big data should play a much bigger role in AD research, especially in areas where subject recruitment and retention are major issues or the time it takes for clinical research results to be generated, e.g., studies involving minorities, epidemiological studies, translation of research findings into practice which average 17 years. The use of EHR could drastically improve the extent, volume, and findings of clinically relevant issues in AD and health care efficiency and outcomes in AD.

The strength of this review includes being the first to analyze the current science of big data, and clinically-relevant research in AD using the most recent studies. One limitation of this review is the difficulty to compare performances of the methods due to studies' variability of data sources and variables selection. We only limited our search to the most recent 5 years and are unable to include all historical details in the field of big data in AD research. Moreover, it's possible that we could have missed some research area foci due to our literature search criteria. We intentionally excluded the literature on bioinformatics such as genome-wide studies and non-clinically related papers.

6. Conclusion

Healthcare data analysis in AD research has driven six research foci using a variety of data sources and data analytics. While the emerging findings are promising, the heterogeneity of study methods hinders the translation and application of the research findings into practice. Future research is needed to generate new hypotheses and replicate existing findings, and fully explore the potentials of EHR data for guiding clinical research and practice.

Acknowledgments

The data analysis and manuscript preparation were supported by the National Center for Complementary and Integrative Health of the National Institutes of Health (NIH) under award number R01AT009457, the National Institute of General Medical Sciences of the NIH award number R01GM120079, and the National Institute on Aging of the NIH award number R01AG043392. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Alzheimer's Disease International. Dementia statistics. 2015. Available at: http://www.alz.co.uk/research/statistics
2. Alzheimer's Association. 2015 Alzheimer's disease facts and figures. Alzheimers Dement. 2015 Mar; 11(3):332–384. [PubMed: 25984581]
3. Cummings JL, Morstorf T, Zhong K. Alzheimer's disease drug-development pipeline: few candidates, frequent failures. Alzheimer's Research & Therapy. 2014; 6(37)
4. IHTT. Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry. Available at: http://ihealthtran.com/wordpress/2013/03/iht%C2%B2-releases-big-data-research-report-download-today/
5. Zhao D, Weng C. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. J Biomed Inform. 2011 Oct; 44(5):859–868. [PubMed: 21642013]
6. Zhang W, Ota T, Shridhar V, Chien J, Wu B, Kuang R. Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. PLoS Comput Biol. 2013; 9(3):e1002975. [PubMed: 23555212]
7. Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN, Miranda ML, et al. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc. 2013 Dec; 20(e2):e319–26. [PubMed: 24026307]
8. Simon GJ, Schrom J, Castro MR, Li PW, Caraballo PJ. Survival association rule mining towards type 2 diabetes risk assessment. AMIA Annu Symp Proc. 2013 Nov 16.2013:1293–1302. [PubMed: 24551408]
9. Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and Machine Learning for Heart Failure Survival Analysis. Stud Health Technol Inform. 2015; 216:40–44. [PubMed: 26262006]
10. Laney D. 3D data management: Controlling data volume, velocity and variety. 2001
11. Cottle M, Hoover W, Kanwal S, Kohn M, Strome T, Treister NT. Transforming Health Care through Big Data.
12. Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC, Yang G. Big data for health. IEEE Journal of Biomedical and Health Informatics. 2015:1193–1208. [PubMed: 26173222]
13. Ross MK, Wei W, Ohno-Machado L. "Big Data" and the Electronic Health Record. Yearb Med Inform. 2014; 9(1):97–104. [PubMed: 25123728]
14. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CRJ, Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's dement. 2011 May; 7(3):263–269. [PubMed: 21514250]
15. Alzheimer's Disease International. World Alzheimer Report 2014 Dementia and Risk Reduction: An Analysis of Protective and Modifiable Factors. 2014. Available at: http://www.alz.co.uk/research/WorldAlzheimerReport2014.pdf
16. Smith PJ, Blumenthal JA, Hoffman BM, Cooper H, Strauman TA, Welsh-Bohmer K, et al. Aerobic exercise and neurocognitive performance: a meta-analytic review of randomized controlled trials. Psychosom Med. 2010 Apr; 72(3):239–252. [PubMed: 20223924]
17. Yu F, Bronas UG, Konety S, Nelson NW, Dysken M, Jack C Jr, et al. Effects of aerobic exercise on cognition and hippocampal volume in Alzheimer's disease: study protocol of a randomized controlled trial (The FIT-AD trial). Trials [Electronic Resource]. 2014; 15:394.
18. Jack CR Jr, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, et al. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011 May; 7(3):257–262. [PubMed: 21514247]
19. Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's dement. 2011 May; 7(3):270–279. [PubMed: 21514249]

20. Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's dement. 2011 May; 7(3):280–292. [PubMed: 21514248]
21. Alzheimer's Disease Big Data DREAM Challenge. Available at: https://www.synapse.org/#!Synapse:syn2290704/wiki/60828
22. Hendrix JA, Finger B, Weiner MW, Frisoni GB, Iwatsubo T, Rowe CC, et al. The Worldwide Alzheimer's Disease Neuroimaging Initiative: An update. Alzheimers Dement. 2015 Jul; 11(7): 850–859. [PubMed: 26194318]
23. van Gils M, Koikkalainen J, Mattila J, Herukka S, Lotjonen J, Soininen H. Discovery and use of efficient biomarkers for objective disease state assessment in Alzheimer's disease. Conf Proc IEEE Eng Med Biol Soc. 2010; 2010:2886–2889. [PubMed: 21095977]
24. Li M, Oishi K, He X, Qin Y, Gao F, Mori S, et al. An efficient approach for differentiating Alzheimer's disease from normal elderly based on multicenter MRI using gray-level invariant features. PLoS One. 2014 Aug 20.9(8):e105563. [PubMed: 25140532]
25. Yang ST, Lee JD, Chang TC, Huang CH, Wang JJ, Hsu WC, et al. Discrimination between Alzheimer's disease and mild cognitive impairment using SOM and PSO-SVM. Comput Math Methods Med. 2013; 2013:253670. [PubMed: 23737859]
26. Mangialasche F, Westman E, Kivipelto M, Muehlboeck JS, Cecchetti R, Baglioni M, et al. Classification and prediction of clinical diagnosis of Alzheimer's disease based on MRI and plasma measures of alpha-/gamma-tocotrienols and gamma-tocopherol. J Intern Med. 2013 Jun; 273(6):602–621. [PubMed: 23343471]
27. Kohannim O, Hua X, Hibar DP, Lee S, Chou YY, Toga AW, et al. Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiol Aging. 2010 Aug; 31(8):1429–1442. [PubMed: 20541286]
28. Clark DG, Kapur P, Geldmacher DS, Brockington JC, Harrell L, DeRamus TP, et al. Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease. Cortex. 2014 Jun.55:202–218. [PubMed: 24556551]
29. Lagun D, Manzanares C, Zola SM, Buffalo EA, Agichtein E. Detecting cognitive impairment by eye movement analysis using automatic classification algorithms. J Neurosci Methods. 2011 Sep 30; 201(1):196–203. [PubMed: 21801750]
30. Casanova R, Hsu FC, Sink KM, Rapp SR, Williamson JD, Resnick SM, et al. Alzheimer's disease risk assessment using large-scale machine learning methods. PLoS One. 2013 Nov 8.8(11):e77949. [PubMed: 24250789]
31. He Z, Wang S, Borhanian E, Weng C. Assessing the collective population representativeness of related type 2 diabetes trials by combining public data from ClinicalTrials.gov and NHANES. Stud Health Technol Inform. 2015; 216:569–573. [PubMed: 26262115]
32. Zhang D, Shen D. Alzheimer's Disease Neuroimaging Initiative. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage. 2012 Jan 16; 59(2):895–907. [PubMed: 21992749]
33. Mattila J, Koikkalainen J, Virkki A, Simonsen A, van Gils M, Waldemar G, et al. A disease state fingerprint for evaluation of Alzheimer's disease. J Alzheimers Dis. 2011; 27(1):163–176. [PubMed: 21799247]
34. Lobo A, Saz P, Día JL, et al. and the ZARADEMP Workgroup. The ZARADEMP Project on the incidence, prevalence and risk factors of dementia (and depression) in the elderly community. II. Methods and first results. Eur J Psychiatry. 2005; 19(1):40–54.
35. Chen R, Young K, Chao LL, Miller B, Yaffe K, Weiner MW, et al. Prediction of conversion from mild cognitive impairment to Alzheimer disease based on bayesian data mining with ensemble learning. Neuroradiol J. 2012 Mar; 25(1):5–16. [PubMed: 24028870]
36. Gracia-Garcia P, de-la-Camara C, Santabarbara J, Lopez-Anton R, Quintanilla MA, Ventura T, et al. Depression and incident Alzheimer disease: the impact of disease severity. Am J Geriatr Psychiatry. 2015 Feb; 23(2):119–129. [PubMed: 23791538]

37. Liu X, Tosun D, Weiner MW, Schuff N. Alzheimer's Disease Neuroimaging Initiative. Locally linear embedding (LLE) for MRI based Alzheimer's disease classification. Neuroimage. 2013 Dec.83:148–157. [PubMed: 23792982]
38. Hinrichs C, Singh V, Xu G, Johnson S. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. NeuroImage. 2011; 55(2):574. [PubMed: 21146621]
39. Challis E, Hurley P, Serra L, Bozzali M, Oliver S, Cercignani M. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. Neuroimage. 2015 May 15.112:232–243. [PubMed: 25731993]
40. Li R, Zhang W, Suk HI, Wang L, Li J, Shen D, et al. Deep learning based imaging data completion for improved brain disease diagnosis. Med Image Comput Comput Assist Interv. 2014; 17(Pt 3): 305–312. [PubMed: 25320813]
41. Lovestone S, Francis P, Strandgaard K. Biomarkers for disease modification trials--the innovative medicines initiative and AddNeuroMed. J Nutr Health Aging. 2007 Jul-Aug;11(4):359–361. [PubMed: 17653500]
42. Costafreda SG, Dinov ID, Tu Z, Shi Y, Liu CY, Kloszewska I, et al. Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. Neuroimage. 2011 May 1; 56(1):212–219. [PubMed: 21272654]
43. Poil SS, de Haan W, van der Flier WM, Mansvelder HD, Scheltens P, Linkenkaer-Hansen K. Integrative EEG biomarkers predict progression to Alzheimer's disease at the MCI stage. Front Aging Neurosci. 2013 Oct 3.5:58. [PubMed: 24106478]
44. Pozueta A, Rodriguez-Rodriguez E, Vazquez-Higuera JL, Mateo I, Sanchez-Juan P, Gonzalez-Perez S, et al. Detection of early Alzheimer's disease in MCI patients by the combination of MMSE and an episodic memory test. BMC Neurol. 2011 Jun 24.11 78-2377-11-78.
45. Defrancesco M, Egger K, Marksteiner J, Esterhammer R, Hinterhuber H, Deisenhammer EA, et al. Changes in white matter integrity before conversion from mild cognitive impairment to Alzheimer's disease. PLoS One. 2014 Aug 25.9(8):e106062. [PubMed: 25153085]
46. Gomar JJ, Conejero-Goldberg C, Davies P, Goldberg TE. Alzheimer's Disease Neuroimaging Initiative. Extension and refinement of the predictive value of different classes of markers in ADNI: four-year follow-up data. Alzheimers Dement. 2014 Nov; 10(6):704–712. [PubMed: 24613706]
47. Alegret M, Cuberas-Borros G, Espinosa A, Valero S, Hernandez I, Ruiz A, et al. Cognitive, genetic, and brain perfusion factors associated with four year incidence of Alzheimer's disease from mild cognitive impairment. J Alzheimers Dis. 2014; 41(3):739–748. [PubMed: 24685632]
48. Runtti H, Mattila J, van Gils M, Koikkalainen J, Soininen H, Lotjonen J, et al. Quantitative evaluation of disease progression in a longitudinal mild cognitive impairment cohort. J Alzheimers Dis. 2014; 39(1):49–61. [PubMed: 24121959]
49. Sager MA, Hermann B, La Rue A. Middle-aged children of persons with Alzheimer's disease: APOE genotypes and cognitive function in the Wisconsin Registry for Alzheimer's Prevention. J Geriatr Psychiatry Neurol. 2005 Dec; 18(4):245–249. [PubMed: 16306248]
50. Li L, Ruau D, Chen R, Weber S, Butte AJ. Systematic identification of risk factors for Alzheimer's disease through shared genetic architecture and electronic medical records. Pac Symp Biocomput. 2013:224–235. [PubMed: 23424127]
51. Chang TS, Coen MH, La Rue A, Jonaitis E, Koscik RL, Hermann B, et al. Machine learning amplifies the effect of parental family history of Alzheimer's disease on list learning strategy. J Int Neuropsychol Soc. 2012 May; 18(3):428–439. [PubMed: 22321601]
52. Rosenberg PB, Mielke MM, Appleby BS, Oh ES, Geda YE, Lyketsos CG. The association of neuropsychiatric symptoms in MCI with incident dementia and Alzheimer disease. Am J Geriatr Psychiatry. 2013 Jul; 21(7):685–695. [PubMed: 23567400]
53. Yasar S, Xia J, Yao W, Furberg CD, Xue QL, Mercado CI, et al. Antihypertensive drugs decrease risk of Alzheimer disease: Ginkgo Evaluation of Memory Study. Neurology. 2013 Sep 3; 81(10): 896–903. [PubMed: 23911756]

54. Malhotra A, Younesi E, Gurulingappa H, Hofmann-Apitius M. 'HypothesisFinder:' a strategy for the detection of speculative statements in scientific text. PLoS Comput Biol. 2013; 9(7):e1003117. [PubMed: 23935466]
55. Greco I, Day N, Riddoch-Contreras J, Reed J, Soininen H, Kloszewska I, et al. Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation. J Transl Med. 2012 Oct 31.10 217-5876-10-217.
56. Li J, Zhu X, Chen JY. Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. PLoS Comput Biol. 2009 Jul.5(7):e1000450. [PubMed: 19649302]
57. Jonnalagadda SR, Del Fiol G, Medlin R, Weir C, Fiszman M, Mostafa J, et al. Automatically extracting sentences from Medline citations to support clinicians' information needs. J Am Med Inform Assoc. 2013 Sep-Oct;20(5):995–1000. [PubMed: 23100128]
58. Parikh M, Hynan LS, Weiner MF, Lacritz L, Ringe W, Cullum CM. Single neuropsychological test scores associated with rate of cognitive decline in early Alzheimer disease. Clin Neuropsychol. 2014; 28(6):926–940. [PubMed: 25131004]
59. Stonnington CM, Chu C, Kloppel S, Jack CR Jr, Ashburner J, Frackowiak RS, et al. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage. 2010 Jul 15; 51(4):1405–1413. [PubMed: 20347044]
60. Chen L, Reed C, Happich M, Nyhuis A, Lenox-Smith A. Health care resource utilisation in primary care prior to and after a diagnosis of Alzheimer's disease: a retrospective, matched case-control study in the United Kingdom. BMC Geriatr. 2014 Jun 17.14 76-2318-14-76.
61. Rattinger GB, Mullins CD, Zuckerman IH, Onukwugha E, Delisle S. Clinic visits and prescribing patterns among Veterans Affairs Maryland Health Care System dementia patients. J Nutr Health Aging. 2010 Oct; 14(8):677–683. [PubMed: 20922345]
62. Baker NL, Cook MN, Arrighi HM, Bullock R. Hip fracture risk and subsequent mortality among Alzheimer's disease patients in the United Kingdom, 1988–2007. Age Ageing. 2011 Jan; 40(1): 49–54. [PubMed: 21087990]
63. Roberts RO, Geda YE, Knopman DS, Cha RH, Pankratz VS, Boeve BF, Ivnik RJ, Tangalos EG, Petersen RC, Rocca WA. The Mayo Clinic Study of Aging: design and sampling, participation, baseline measures and sample characteristics. Neuroepidemiology. 2008; 30(1):58–69. [PubMed: 18259084]
64. Obermeyer Z, Emanuel J. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016; 375:1216–1219. [PubMed: 27682033]
65. Chen J, Asch S. Machine Learning and Prediction in Medicine – Beyond the Peak of Inflated Expectations. N Engl J Med. 2017; 376:2507–2509. [PubMed: 28657867]
66. Chao CA. The impact of electronic health records on collaborative work routines: A narrative network analysis. Int J Med Inform. 2016 Oct.94:100–11. [PubMed: 27573317]
67. Weng C. Optimizing Clinical Research Participant Selection with Informatics. Trends Pharmacol Sci. 2015 Nov; 36(11):706–9. [PubMed: 26549161]
68. Vemuri P, et al. Vascular and amyloid pathologies are independent predictors of cognitive decline in normal elderly. Brain. 2015; 138(3):761–771. [PubMed: 25595145]
69. Vemuri P, et al. Cognitive reserve and Alzheimer's disease biomarkers are independent determinants of cognition. Brain. 2011; 134(5):1479–1492. [PubMed: 21478184]
70. Amieva H, et al. Compensatory mechanisms in higher-educated subjects with Alzheimer's disease: a study of 20 years of cognitive decline. Brain. 2014; 137(4):1167–1175. [PubMed: 24578544]
71. Stern Y. Cognitive reserve in ageing and Alzheimer's disease. Lancet Neurol. 2012; 11(11):1006–1112. [PubMed: 23079557]

Summary points

What was already known on the topic

- Alzheimer's disease (AD) is the most common type of dementia constituting 60–80% of all dementias.
- AD is the only disease that cannot be prevented, slowed, or cured.

What this study added to our knowledge

- Big data research in AD is growing in the recent years
- Big data research in AD mainly address the six research foci: AD or MCI diagnosis, prediction of MCI to AD conversion, stratification of AD risks, knowledge discovery from literature, prediction of AD progression, and description of clinical care for persons with AD.
- Majority of big data research in AD used the existing research databases, including Alzheimer's Disease Neuroimaging Initiative database and AddNeuroMed study.
- EHR provides big data resource to potentially support AD clinical research.

Highlights

- Big data are important to advance research in Alzheimer's disease (AD) due to the difficulties in recruitment and retention of patients in clinical research and the durations and costs associated with traditional clinical research.
- We analyzed 38 studies to derive 7 research foci inductively, including diagnosing AD or mild cognitive impairment (MCI), predicting MCI to AD conversion, stratifying risks for AD, mining the literature for knowledge discovery, predicting AD progression, describing clinical care for persons with AD, and understanding the relationship between cognition and AD.
- The datasets used for AD research include Alzheimer's Disease Neuroimaging Initiative (ADNI), electronic health records (EHR), MEDLINE, and other research datasets.
- Data analytics methods cover a wide range including data mining, machine learning, natural language processing (NLP), text mining and statistical analysis.
- Big data in AD research is still in its early stage and more efforts should integrate real world big data to advance AD research and practice.

Table 1

Selected data sources for AD research.

Data source	Population	Number of patients	Kind of data elements
Alzheimer's Disease Neuroimaging Initiative (ADNI)	59 sites in USA	• ADNI1: 200 AD, 400 MCI, 200 HC • ADNI GO: 200 EMCI, 500 HC, and 400 MCI (from ADNI1) • ADNI2: 200 AD, 150 LMCI, 350 EMCI (200 from ADNI GO), 450– 500 CN and MCI (from ADNI1) and 150 HC	Demographic, imaging, genetics, clinical data
AddNeuroMed study	6 sites in Europe	378 subjects (130 AD, 131 MCI, 117 HC)	Demographic, imaging
Mayo Clinic Study of Aging	Olmsted County, Minnesota, USA	2719 subjects (402 dementia)	Demographic, imaging, genetics, clinical data
ZARAgoza DEMentia DEPression (ZARADEMP) study	Zaragoza, Spain	• Zarademp I: 4803 participants • Zarademp II: 3237 participants • Zarademp III: 2403 participants	Demographic, cognitive measures, physical measures, questionnaires
Electronic Health Records	Clinical practice research datalink, UK	3896 AD, 7792 HC	Medical resources utilization (consultation, specialty referral, length of hospitalization)

Table 2

Search terms used in literature search.

Category	Query terms
AD	Alzheimer Disease
	Alzheimer's Disease
AND	
Data	Big Data
	Clinical Data
	Health Data
	Healthcare Data
	Electronic Health Records
	Electronic Medical Records
Data Analysis	Data Analysis
	Data Mining
	Machine Learning
	Predictive Model
	Text Mining
	Natural Language Processing

Table 3

Datasets, data analysis methods, and key findings for selected studies in research foci 1 - diagnosing AD or MCI.

Study	Datasets	Primary data analysis methods	Key findings
Van Gils et al. (23)	ADNI data (229 HC, 402 MCI, 190 AD) and data from Kuopio L-MCI study (687 HC, 249 MCI, 77 AD)	Support vector machine (SVM) and linear model. Performance of separating persons with AD from those with MCI or HC using combined features is 94%–100%.	Identification of efficient biomarker sets (including Apolipoprotein E (ApoE) alleles, CSF, estrogen usage duration, cognitive and memory tests and MRI features) related to AD diagnosis
Li et al. (24)	MRIs from ADNI1 (80 AD, 141 MCI, 142 HC)	Support vector machine with selected imaging features from structured MRI.	Selected AD-specific anatomical features from structured MRI have discriminative capability in differentiating AD or MCI from healthy controls.
Yang et al. (25)	Clinical dementia rating, MMSE and MRI scans from 17 AD, 18 MCI, 17 HC	Support vector machine with particle swarm optimization (PSO) and principle component analysis (PCA). Diagnosis accuracy: 94% for AD and 88.9% for MCI	SVM-PSO with PCA can classify AD and MCI versus HC
Mangialasche et al. (26)	Structural MRI measures, plasma levels of vitamin E and makers of vitamin E oxidative/nitrosative damage (81 AD, 86 MCI and 86 HC from AddNeuroMed study)	Multivariate data analysis (Orthogonal partial least squares to latent structures (OPLS)), with 67 variables from structural MRI measures and plasma levels of vitamin E forms.	Plasma levels of tocopherols and tocotrienols with MRT can differentiate AD and MCI from HC subjects and predict MCI to AD conversion
Kohannim et al. (27)	MRI and biomarkers from ADNI (158 AD, 366 MCI, 213 HC)	Support vector machine with brain imaging and other biomarkers features	SVM with brain imaging and biomarkers can classify AD, MCI and HC.
Clark et al. (28)	Semantic fluency word lists, dementia rating, and neuropsychological assessment (training set: 41 AD, 80 MCI, 44 HC; testing set: 9 AD, 21 MCI, 35 HC)	Random forest classifier	Semantic fluency lists can potentially predict functional declines
Lagun et al. (29)	Eye movement data and neuropsychological assessment (20 AD, 10 MCI, 30 HC)	Naïve Bayes, Logistic regression, Support vector machine	Eye movement measures with SVM classification techniques can detect MCI
Casanova et al. (30)	Structural MRI, DNA, and cognitive data of Caucasians in ADNI (171 AD, 153 PMCI, 182 SMCI, 188 CN)	Regularized logistic regression (RLR)	A new metrics, AD pattern similarity (AD-PS) scores, was designed to assess risk of AD.
Zhang et al. (32)	MRI, FDG-PET, CSF data from ADNI (45 AD, 91 MCI, 50 HC)	Multi-modal multi-task (M3T) learning	M3T learning performed well on both AD detection and clinical score prediction
Mattila et al. (33)	ADNI (163 AD, 190 SMCI, 154 PMCI, 199 HC)	Disease state fingerprint visualization, statistical disease state index (DSI) method	DSI can estimate AD state

Note: ADNI, Alzheimer's Disease Neuroimaging Initiative database; MCI, mild-cognitive impairment; SMCI, stable mild cognitive impairment; PMCI, progressive mild cognitive impairment; HC, healthy control.

Table 4

Datasets, data analysis methods, and key findings for selected studies in research foci 2 - predicting MCI to AD conversion.

Study	Datasets	Primary data analysis methods	Key findings
Chen et al. (35)	MRI and magnetic resonance spectroscopy data (8 PMCI, 18 SMCI), ADNI (for validation)	Bayesian network (Bayesian Outcome prediction with Ensemble Learning (BOPEL)) based on brain volums in different regions	Bayesian data mining with ensemble learning demonstrates high predictive accuracy for MCI to AD conversion.
Gracia- Garcia et al. (34,36)	3,864 participants from a longitudinal epidemiological study Zaragoza Dementia and Depression Project (ZARADEMP)	Multivariate model	Severe depression increases the risk of AD
Liu et al. (37)	MRI data from ADNI (86 AD, 93 SMCI, 97 MCI converters, 137 HC)	Locally linear embedding (LLE) by transforming multivariate MRI data to a locally linear space	LLE can improve the performance for predicting MCI to AD conversion
Hinrichs et al. (38)	ADNI subjects with MRI and FDG-PET scans (48 AD, 119 MCI, 66HC)	Support vector machine, Multi-kernel learning	MKL outperforms SVM. Imaging modalities are better predictors than neuropsychological scores.
Challis et al. (39)	Clinical measurements and MRI scans from 27 AD, 50 MCI and 39 HC	Bayesian Gaussian process logistic regression (GP-LR)	GP-LR can accurately differentiate MCI versus HC and MCI versus AD.
Li et al. (40)	MRI scans from ADNI database (180 AD, 160 MCI converters, 214 MCI non-converters, 204 HC)	Deep learning	Deep learning can distinguish four stages of AD progression using MRI with less clinical prior knowledge.
Lovestone et al. (41,42)	Behavioral assessment, hippocampal morphology, MRI in AddNeuroMed study (71 AD, 103 MCI, 88 HC)	Support vector machine	Hippocampal shape analysis provides a prognostic biomarker to predict MCI to AD conversion
Poil et al. (43)	MRI, EEG and clinical data (25 PMCI, 39 SMCI) from subjects referred to Alzheimer Center in Netherland	Elastic net logistic regression	Six EEG biomarkers mainly related to activity in the beta-frequency range (13–30 Hz) can predict conversion from MCI to AD (sensitivity of 88%, specificity of 82%)
Pozueta et al. (44)	Clinical and neuropsychological evaluation (55 SMCI, 50 PMCI)	Logistic regression	A combination of MMSE and an episodic memory test can predict MCI to AD conversion. (PPV: 80.95%)
Zhang et al. (32)	MRI, FDG-PET, CSF data from ADNI (45 AD, 91 MCI, 50 HC)	Multi-modal multi-task (M3T) learning	M3T learning performed well on both AD detection and clinical score prediction
Gomar et al. (46)	Clinical, cognitive, MRI, positron emission tomography, and cerebrospinal fluid from ADNI1 (150 PMCI, 168 SMCI)	Logistic regression	Cognitive measures especially an episodic memory measure and clock drawing screening test were evaluated to be great predictors for MCI to AD conversion.
Alegret et al. (47)	Neuropsychological measurements, brain SPECT data from 42 AD, 42 MCI (25 PMCI, 14 SMCI), and 42 HC	Correlation analysis, Cox regression analysis	Extent of memory impairment associated with speed of MCI to AD conversion
Runtti et al. (48)	Neuropsychological tests and AD biomarkers in ADNI (140 PMCI, 149 SMCI)	Machine learning based disease state index (DSI) method, linear regression	DSI to quantify longitudinal clinical data can predict MCI to AD conversion. (76.9% accuracy)

Note: ADNI, Alzheimer's Disease Neuroimaging Initiative database; MCI, mild-cognitive impairment; SMCI, stable mild cognitive impairment; PMCI, progressive mild cognitive impairment; HC, healthy control.

Table 5

Datasets, data analysis methods, and key findings for selected studies in research foci 3 - stratifying risks for AD.

Study	Datasets	Primary data analysis methods	Key findings
Gracia- Garcia et al. (34,36)	3,864 participants from ZARADEMP	Multivariate model	Severe depression increases the risk of AD
Li et al. (50)	Clinical data from EHR (212 AD, 15040 HC)	Statistical analysis (Chi-square test and Mann-Whitney U test)	Erythrocyte sedimentation rate (ESR) is a significantly associated with AD.
Chang et al. (51)	879 asymptomatic higher risk persons (with parental family history of AD) and 355 asymptomatic lower risk persons (without parental family history of AD) from WRAP	Aggregate measure using Euclidean distance	Finer differences in memory strategy measured by machine learning method can be used as a potential AD risk factor.
Rosenberg et al. (52)	1821 MCI (527 PMCI, 454 SMCI) from National Alzheimer's Coordinating Center database	Cox proportionality hazard model	Neuropsychiatric symptoms in MCI are associated with significantly increase of incident dementia and AD.
Yasar et al. (53)	320 MCI and 1928 HC from Ginkgo Evaluation of Memory study	Cox proportional hazard model	The use of diuretic, angiotensin-1 receptor blocker (ARB), and angiotensin-converting enzyme inhibitors (ACE-I) was associated with reduced AD risk for healthy normal. Diuretic use associated with reduced AD risk in MCI patients.

Note: ZARADEMP, Zaragoza Dementia and Depression Project; WRAP, Wisconsin Registry for Alzheimer's Prevention; MCI, mild-cognitive impairment; SMCI, stable mild cognitive impairment; PMCI, progressive mild cognitive impairment; HC, healthy control.

Table 6

Datasets, data analysis methods, and key findings for selected studies in research foci 4 - mining the literature for knowledge discovery.

Study	Datasets	Primary data analysis methods	Key findings
Malhotra et al. (54)	3,007 MEDLINE abstracts, and annotations of non-overlapping 200 randomly chosen abstracts	HypothesisFinder (including dictionary based named entity recognition and rule-based pattern recognition)	HypothesisFinder can build hypothetical protein interaction network.
Greco et al. (55)	Full text literature reviews, MEDLINE, web-based reports, and databases (e.g., gene expression databases, protein-pathway databases, protein-disease association databases)	Text mining	25 potential candidate biomarkers were found. Two candidate biomarkers (i.e., Choline Acetyltransferase (ChAt) and urokinase-type Plasminogen Activator Receptor (PLAUR)) have not been reported previously.
Li et al. (56)	Online Predicted Human Interaction Database (OPHID) for protein interaction, 222,609 AD-related MEDLINE abstracts for AD drugs	Molecular interaction network mining, text mining	17 candidate AD drugs were found
Jonnalagadda et al. (57)	MEDLINE abstracts	Information extraction and semantic information extraction	336 AD-related sentences from 194 abstracts in the MEDLINE, 84.5% of which were evaluated as relevant abstracts

Table 7

Datasets, data analysis methods, and key findings for selected studies in research foci 5 - predicting AD progression.

Study	Datasets	Primary data analysis methods	Key findings
Parikh et al. (58)	Neuropsychological, clinical and psychiatric measures for 96 patients with mild AD (45 faster and 51 slower progresses) enrolled at AD Center at the University of Texas Southwestern Medical Center during 1995–2011	Stepwise logistic regression	Several neuropsychological performance features can predict cognitive decline rate in mild AD.
Stonnington et al. (59)	MRI scans in ADNI (113 AD, 351 MCI, 122 HC) and MMSE, dementia rating scale (DRS), Auditory Verbal Learning Test (AVLT) measures from Mayo clinic (73 AD and 91 HC)	Relevance vector regression (RVR)	RVR can predict MMSE, Dementia Rating Scale (DRS) and Alzheimer's Disease Assessment Scale—Cognitive subtest ADAS and can aid in AD diagnosis and clinical outcome prediction.

Note: ADNI, Alzheimer's Disease Neuroimaging Initiative database; MCI, mild-cognitive impairment; SMCI, stable mild cognitive impairment; PMCI, progressive mild cognitive impairment; HC, healthy control.

Table 8

Datasets, data analysis methods, and key findings for selected studies in research foci 6 - describing clinical care for persons with AD.

Study	Datasets	Primary data analysis methods	Key findings
Chen et al. (60)	EHR data from UK's Clinical Practice Research Datalink (3896 AD, 7792 HC)	Generalized linear model	AD diagnosis associated with significant increase in primary and secondary care recourse utilization
Baker et al. (62)	EMR data in UK (10,052 AD, 10,052 HC)	Survival curve, Cox regression analysis	AD patients had a higher incidence of hip fracture.
Rattinger et al. (61)	EMR data from Veterans Affairs Maryland Health Care System (1209 dementia patients)	Logistic regression	Visits to specialized dementia or mental health clinical increases the odds of receiving anti-demantic, antidepressant and antipsuchotic medication

Table 9

Datasets, data analysis methods, and key findings for selected studies in research foci 7 – understanding the relationship between cognition and AD.

Study	Datasets	Primary data analysis methods	Key findings
Vemuri et al. (68)	Clinical and cognitive assessments, CSF, MRI (98 AD, 192 MCI, 109 HC from ADNI)	Correlation analysis	Cognitive reserve and biomarkers of AD pathology are independent predictors of cognitive performance
Vemuri et al. (69)	Imaging biomarkers, cognitive reserve variables, global cognition measure (393 healthy subjects with one clinical follow-up from MCSA)	Association analysis, linear mixed model	Cognitive reserve offset the deleterious effect of pathologies of the cognitive trajectories
Amieva et al (70)	Neuropsychological evaluation from 442 AD subjects with 20 years follow-up (171 low education, 271 high education) from PAQUID cohort	Linear mixed model	Cognitive decline occurred up to 16 years and 7 years before AD for high- educated and low-educated individuals, respectively