Semin Speech Lang 2016; 37(01): 003-009
DOI: 10.1055/s-0036-1572385
Thieme Medical Publishers 333 Seventh Avenue, New York, NY 10001, USA.

The Rise of Big Data in Neurorehabilitation

Yasmeen Faroqi-Shah[1]

Department of Hearing and Speech Sciences, University of Maryland, College Park, Maryland

Abstract

In some fields, Big Data has been instrumental in analyzing, predicting, and influencing human behavior. However, Big Data approaches have so far been less central in speech-language pathology. This article introduces the concept of Big Data and provides examples of Big Data initiatives pertaining to adult neurorehabilitation. It also discusses the potential theoretical and clinical contributions that Big Data can make. The article also recognizes some impediments in building and using Big Data for scientific and clinical inquiry.

Keywords

Big Data - open science - TalkBank - AphasiaBank - Alzheimer disease - aphasia
Learning Outcomes: As a result of this activity, the reader will be able to (1) describe the concept of Big Data and its relevance to adult neurogenic disorders; (2) identify examples of Big Data that pertain to adult neurogenic disorders; (3) discuss the potential and challenges of using Big Data to advance the science and clinical practice of adult neurogenic disorders.

The term *Big Data* is ubiquitous today, and it represents different ideas to different people. It refers to enormous quantities of diverse data that cannot be handled by traditional data management systems. Such data could be generated by electronic record keeping (such as retail sales and patient medical records), human activities (such as Web browsing and cell phone use), or by machines (such as weather radar). Other Big Data sources are incidental by-products of human activity; still others are intentionally, and often collaboratively, assembled with an end goal. These voluminous data sets can be unstructured stockpiles accrued over time, or fairly well structured (such as diagnosis-coded Medicare data). Thus, to some, *Big Data* also refers to ways of extracting behavior patterns from unstructured data, such as social media analytics. Additionally, it refers to next-generation data management capabilities that can handle increasingly large volumes of data. Although some Big Data repositories have limited accessibility, others are openly available for public data mining.

Regardless of one's specific definition of Big Data, it is undeniable that various disciplines and organizations are beginning to recognize the importance of Big Data in analyzing, predicting, and influencing human behavior. *Open Science*, which refers to making scientific research and data accessible to everyone, when combined with Big Data, has immense potential for innovation and scientific development. The goal of this issue of *Seminars in Speech and Language* is to highlight the application of Big Data approaches to the science and clinical practice of adult neurorehabilitation. This article introduces the topic by first presenting an overview of relevant Big Data sources in adult neurorehabilitation. This will be followed by a discussion of opportunities and challenges in using Big Data and potential clinical applications.

Big Data and Adult Neurogenic Disorders

Big Data is not exactly new in language and neurobehavioral research. The best illustration of the immense potential of Big Data in scientific investigations of adult neurogenic disorders comes from Alzheimer disease (AD), which is among the most prevalent and rapidly growing diseases in the elderly population.[1] There are multiple repositories of AD data; one of these is the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS),[2] which was established by the National Institute of Aging in collaboration with the University of Pennsylvania as a data repository for research funded by the National Institute of Aging. NIAGADS currently has genomic data of over 49,000 individuals with late-onset AD.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) was launched by a partnership between the National Institute of Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, along with private pharmaceutical companies and nonprofit organizations.[3] ADNI was developed with the goal of tracking the progression of mild cognitive impairment and early AD using magnetic resonance imaging (MRI), positron emission tomography, other biological markers, and clinical and neuropsychological assessment. It has data of over 1,500 individuals. The data in both the NIAGADS and ADNI were built bottom-up from the efforts of many investigators from a broad range of academic institutions and private corporations, and participants have been recruited from over 50 sites. In 2014, such Big Data initiatives were utilized in the Alzheimer's Disease Big Data DREAM Challenge,[4] a competitive effort to identify predictive biomarkers of cognitive decline in AD. Determination of biomarkers of early AD progression will aid researchers and clinicians to identify populations at risk, use preventive care, and develop new and effective treatments. There are also AD Big Data from public health records and other data consortia with multi-investigator contributions.[5] [6] [7] All these AD Big Data efforts have allowed researchers to focus their energies on data mining to better understand the epidemiology of AD, saving the cost of new data collection.

In recognition of the importance of gaining a better understanding of human brain structure and function, there are several Big Data initiatives pertaining to imaging of the human brain. For instance, the Human Connectome Project is an outcome of the National Institutes of Health Blueprint for Neuroscience Research with the goal of mapping neural pathways that underlie human brain function.[8] It aims to obtain high-resolution structural, functional, and diffusion imaging scans from 1,200 healthy individuals, including twins and their nontwin siblings. The data are being collected at two consortium clusters using a standard data collection protocol.

A different approach to Big Data of the human brain is NeuroVault,[9] which is an investigator-initiated international repository of MRI and positron emission tomography scans supported by the International Neuroinformatics Coordination Facility. The goal of NeuroVault is to uncover patterns of (sometimes subtle) brain activity that may be consistently found across studies but are not reported in published research. Such studies may not report certain brain activations because either the activity or brain region was not the focus of the study or the anomalies did not meet statistical threshold. NeuroVault currently contains unthresholded statistical maps of over 240 published studies uploaded by researchers.

Yet another approach to uncovering consistencies of human brain function across studies is utilized by Neurosynth,[10] which was developed by team members of the NeuroVault. In the Neurosynth platform, published neuroimaging articles are automatically parsed to extract locations of functional brain activations and key terms used in the article full text (e.g., language, working memory). The platform then performs a meta-analysis of studies that contain the key term. Neurosynth currently has images from over 11,000 brain imaging studies, and the automated meta-analyses can be easily accessed by anyone.

A similar, but statistically more conservative, meta-analytic approach called *activation likelihood estimation* is used by BrainMap.[11] This open platform, however, involves manual data entry by investigators wishing to perform a meta-analysis. BrainMap currently has data from over 2,700 published studies covering over 13,000 experiments.

Finally, for the speech-language clinician, user-friendly neuroimaging Big Data platforms such as Neurosynth offer empirically derived current evidence on brain function that can inform predictions and observations about making connections between brain lesion and behavioral deficits in their clinical practice.[10]

Currently there are only a few examples of Big Data in aphasiology. The Moss Aphasia Psycholinguistics Project Database was compiled with patient data from the Moss Rehabilitation and Research Institute in Philadelphia.[12] It contains trial-by-trial picture naming data from over 240 persons with aphasia for the Philadelphia Naming Test, along with other speech, language, and cognitive data, and can be accessed by researchers. This database has been used in several studies that examined the relationship between brain lesions and language symptoms in aphasia.[13] [14] [15]

Whereas the Moss Aphasia Psycholinguistics Project Database primarily contains data from a single data collection site,[12] the AphasiaBank grew (and continues to grow) from contributions by multiple global collaborators, adding languages such as Cantonese, Italian, and Spanish.[16] The database currently contains multimedia discourse samples from over 285 individuals, some from multiple visits to permit longitudinal aphasia data for a total of 408 transcripts of discourse in aphasia. In addition, 180 nonaphasic subjects were also tested and videotaped using the same protocol. The database also contains language test scores of individuals with aphasia (see MacWhinney and Fromm, this issue). The goal of AphasiaBank is to provide researchers and clinicians with a large shared multimedia database of language samples and test scores and to provide automated language analysis tools that facilitate not only scientific investigations but also clinical documentation by speech-language pathologists.[17] AphasiaBank is a part of the TalkBank database,[18] which has discourse samples of other populations including children, bilingual speakers, and individuals with dementia and traumatic brain injury. TalkBank developed from the Child Language Data Exchange System, which was created in the 1980s to collate child language samples.[19] TalkBank is by far the largest collection of discourse samples freely available.

Researchers and clinicians interested in examining patterns of second language speakers of English could use the International Corpus of Learner English (ICLE),[20] which contains English language samples of speakers whose first languages include more than 15 languages such as Bulgarian, Chinese, Czech, and Dutch. The ICLE is the result of a large-scale global collaboration between 19 universities. Similar to the AphasiaBank, language sampling in the ICLE utilized a standard elicitation procedure.

The Big Data for adult neurorehabilitation described so far are primarily of diagnostic and predictive significance. Additionally valuable to clinicians would be large-scale recovery and intervention data. Such data would help determine dose–response relationships with interventions, the relative effectiveness of interventions, and the synergistic effects of different prognostic factors. One such intervention focused data-sharing initiative is the Collaboration of Aphasia Trialists (CATs),[21] which is supported by the European Cooperation in Science and Technology. It consists of a multidisciplinary group of investigators focused on aphasia from rehabilitation, social science, and language research perspectives. The goal of CATs is to create a large network of shared knowledge, skills, and methodology relating to aphasia research that will eventually enhance diagnosis, prognosis, and rehabilitation of aphasia. The CATs currently has more than 600 data sets and several ongoing projects.

In the United States, the American Speech-Language Hearing Association's (ASHA) National Outcome Measurement System (NOMS) is a mechanism that collects data on patient demographics, neurologic and medical history, functional outcome measures, cognitive-linguistic scores, and duration and type of speech-language therapy.[22] The most recent NOMS report for 2006 to 2010 has data for over 6,500, 19,000, 11,000, and 43,000 individuals, respectively, from acute care hospitals, inpatient rehabilitation, and outpatient and skilled nursing settings.[22] The ASHA NOMS data are currently accessible to organizations who participate in the data reporting, and descriptive summary data can be accessed by ASHA members.

A publicly accessible rehabilitation-related infrastructure initiative in the United States is the National Institutes of Health–funded Center for Rehabilitation Research Using Large Datasets,[23] which aims to improve the quantity and quality of large-scale rehabilitation data by providing administrative, training, and infrastructure support to researchers. It involves a consortium of investigators from the University of Texas, Cornell University, and the University of Michigan and contains easily searchable public data sets such as National Nursing Home Survey, conducted by the Centers for Disease Control and Prevention, and the Medicare Health Outcomes Survey, conducted by Department of Health and Human Services. The Center for Rehabilitation Research Using Large Datasets also provides access to disability statistics.

A final Big Data category is emerging from various online platforms used for self-directed learning.[24] For instance, aphasia therapy apps such as Constant Therapy and Talkpath log user performance data.[25] [26] Given the popularity of such platforms, with hundreds of users, these have the potential of increasing our understanding of initial performance, learning outcomes, and prognosis (see Kiran, this issue). Reliability and validity of the Neurocognitive Performance Test, which is a Web-based self-administered cognitive assessment, was recently published using data from over 130,000 individuals.[27] However, most of these app-based data sets are not publicly available. Although the data sets pertaining to adult neurorehabilitation are not (yet) at the huge scale of Big Data in other fields (such as weather patterns or social media use), they represent valuable opportunities for advances in science and clinical practice.

Opportunities of Big Data

One may ask how and why Big Data is relevant for speech-language pathology. Let's illustrate with the example of aphasia. It is well known that language breakdown and its recovery in persons with aphasia is multifaceted,[28] affecting language components (lexical-semantic, syntactic, phonological) and modalities (comprehension, expression, orthography) differently across individuals. Additionally, the neurologic and general health status of the individual influences initial severity and recovery.[28] And individual cognitive, psychoemotional, social, and lifestyle variables further contribute to functional outcomes. This intricate combination of linguistic, health-related, cognitive, and psychosocial variables that determines language breakdown/recovery is incompletely understood, despite decades of scientific inquiry. Investigating the influence of and the interaction between these multiple factors necessitates recruitment of large numbers of individuals with aphasia. A challenge to this endeavor is the inherent variability among persons with aphasia, coupled with small sample sizes in aphasia research. In fact, the average sample size in aphasiological research is 13.2 per study (extracted from articles published in *Aphasiology* 2013). Obstacles to conducting large-group research include limited access to individuals with aphasia in certain geographic regions and organizational settings (e.g., nonhospital), financial and personnel cost, and time constraints. The number of participants is particularly small for intervention research, given the time commitment in administering face-to-face intervention. Small samples of heterogeneous participants stymie researchers' ability to draw statistically reliable and generalizable inferences about language breakdown and recovery. The scenario is similar for other adult neurogenic populations, and some groups, such as those with traumatic brain injury, tend to be even more heterogeneous than individuals with aphasia. Big Data offers opportunities to address this sample size problem in our field.

Big Data initiatives can create a community of researchers who work jointly to address the same research question(s) on the same large data set. Collecting data using a standard protocol accelerates data collection, and openly sharing these data can grow the data set to a size likely unachievable by single or small groups of investigators. Big Data examples discussed earlier, such as ADNI and AphasiaBank, have used this protocol standardization and sharing approach with considerable success. MacWhinney and Fromm (this issue) discuss the various investigations using data from AphasiaBank.

In addition to allowing statistically robust analyses, Big Data offers other opportunities. Data sharing provides access to data for new and seasoned investigators, thus fostering data democratization. It

allows for replication and verification of research across linguistic/cultural and geographic differences. For example, Miyashita and colleagues examined the genetic bases of AD using a genomewide association study in three ethnic populations—Caucasian, Japanese, and Korean—totaling over 11,000 subjects with AD and 10,000 control individuals.[7] This large-scale study was made possible with access to shared Big Data. Importantly, analysis of shared Big Data allows one to evaluate questions blindly without selection bias, and thus contributes to rigorous theory testing.

Needless to say, theoretical understanding of neurologic impairments and their recovery does not automatically emerge from Big Data; it requires a deliberate effort to use the data to extract scientific principles or validate existing theories. An illustration of an open science approach to unbiased theory testing is the Alzheimer's Disease Big Data DREAM Challenge,[4] which sought to uncover the best predictive model for AD. The challenge posed three questions, predicting cognitive scores 24 months after initial assessment and predicting which cognitively normal individuals were likely to develop neuropathological signs of AD and classifying diagnostic groups based on MRI findings. Across these three subchallenges, close to 100 groups of investigators submitted predictive models. Following this, the most accurate predictive models, those that best mimicked actual Big Data, were identified,[3] resulting in a better understanding of AD.

As mentioned earlier, there is a dearth of recovery and outcomes information in adult neurorehabilitation. Although it is challenging to track individuals after they have been discharged from acute and subacute rehabilitation, one could envision a massive shared database of (deidentified) information on patient demographics, neurologic and medical history, functional outcome measures, cognitive-linguistic scores, duration and type of therapy, and most importantly, repeated measures of cognitive-linguistic scores over time and community reintegration information. This would allow researchers to develop predictive models of prognosis and functional outcomes. Although CATs and ASHA's NOMS collect some of these measures,[21] [22] repeated measures of the same individuals' cognitive-linguistic scores over time and long-term community reintegration data would be particularly valuable. These data could allow for powerful statistical modeling of factors that influence prognosis and community reintegration, especially if individuals could be tracked over time both within a setting and as they transition across multiple settings. Augmenting the current information on duration of therapy with types of therapies/goals would be valuable in comparing effectiveness as well as determining evidence-based candidacy for various therapies.

In summary, Big Data, especially when coupled with open science, could unleash innovative approaches to improved diagnostics, prognostics and therapies for adult neurorehabilitation. Indeed, Big Data initiatives have already yielded a better understanding of adult neurogenics.[5] [7] [13] [14] [15] [16] [17] [21] [29] Although Big Data has the potential to advance science and service delivery, there are also challenges in its implementation, which will be discussed next.

Challenges of Big Data

The power of Big Data lies in having a structure that is conducive to data mining; unsystematic piles of Big Data are likely to consume immense effort to sift through and extract reliable patterns. Hence standard data collection protocols are essential, which require collaborative and concerted long-term planning. In the case of international or cross-linguistic repositories, the availability of comparable normed assessment tools might be a limiting factor. Users have to assume that all Big Data contributors have uniformly implemented data-collection protocols, otherwise the integrity of the data and ensuing findings are jeopardized. One risk of open data sharing is ensuring that the privacy of research participants is safeguarded. Furthermore, data may need time-consuming preprocessing before it can be shared or made available for data mining. For example, the discourse samples of AphasiaBank need to be transcribed in CHAT format by the researchers collecting the discourse samples before uploading to AphasiaBank.[19] Depending on the sample length, this could take several hours. Developing a Big Database not only needs collaborations, consistency, privacy safeguards, and a data management plan, it also requires long-term financial support to sustain the

data infrastructure over the long term. Finally, Big Data is only as good as the tools available for analysis. This necessitates development of innovative and efficient data analysis tools, which may require expertise. For instance, the use of TalkBank/AphasiaBank is facilitated by tools such as CLAN and EVAL.[18]

Conclusions

The growing availability of shared databases with contributions from multiple researchers, coupled with high-performance computing, offers opportunities to better understand the multifactorial behavior of neurogenic impairments and address population heterogeneity and the sample size problem in our field. Open data sharing improves accessibility for researchers, thus there is more research because barriers to entry are lowered. A benefit of shared data sets is that the research is better, because results can be replicated and theories can be compared. It creates a community of people among whom ideas and tools circulate rapidly, and it contributes to the growth of the field. Currently, there are a few examples of Big (and not so Big) Data in adult neurorehabilitation; however, there is a vast need for further development of Big Data, and this holds the promise of better elucidating impairment and recovery mechanisms, which would in turn improve clinical service delivery. This issue of *Seminars in Speech and Language* aims to draw attention to the relevance of Big Data to adult neurorehabilitation. The next four articles showcase how analysis of Big Data has tested existing theories and provided novel insights into the intricacies of language breakdown and recovery in aphasia.

However, Big Data initiatives are not without challenges. Some of its proponents believe that ultimately Big Data will create a scientific paradigm shift. However, it seems likely that Big Data applied to scientific exploration will continue to rely on strong behavioral heroes as well as careful and principled experimentation.

References

1 Akushevich I, Kravchenko J, Ukraintseva S, Arbeev K, Yashin AI. Time trends of incidence of age-associated diseases in the US elderly population: Medicare-based analysis. Age Ageing 2013; 42 (4) 494-500

2 National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) . University of Pennsylvania. Available at: https://www.niagads.org/ . Accessed December 12, 2015

3 Weiner MW. Alzheimer's Disease Neuroimaging Initiative (ADNI). Available at: http://adni.loni.ucla.edu/ . Accessed December 12, 2015

4 Alzheimer's Disease Big Data Dream Challenge 1 . Available at: https://www.synapse.org/#!Synapse:syn2290704/wiki/60828 . Accessed January 5, 2016

5 Kuwano R, Miyashita A, Arai H , et al; Japanese Genetic Study Consortium for Alzeheimer's Disease. Dynamin-binding protein gene on chromosome 10q is associated with late-onset Alzheimer's disease. Hum Mol Genet 2006; 15 (13) 2170-2182

6 Beck T, Hastings RK, Gollapudi S, Free RC, Brookes AJ. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. Eur J Hum Genet 2014; 22 (7) 949-952

7 Miyashita A, Koike A, Jun G , et al; Alzheimer Disease Genetics Consortium. SORL1 is genetically associated with late-onset Alzheimer's disease in Japanese, Koreans and Caucasians. PLoS ONE 2013; 8 (4) e58618

8 Human Connectome Project (HCP) . Available at: http://www.humanconnectome.org/ . Accessed December 12, 2015

9 Gorgolewski KJ, Varoquaux G, Rivera G , et al. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Front Neuroinform 2015; 9: 8

10 Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD. Large-scale automated synthesis of human functional neuroimaging data. Nat Methods 2011; 8 (8) 665-670

11 Eickhoff SB, Laird AR, Grefkes C, Wang LE, Zilles K, Fox PT. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. Hum Brain Mapp 2009; 30 (9) 2907-2926

12 Mirman D, Strauss TJ, Brecher A , et al. A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. Cogn Neuropsychol 2010; 27 (6) 495-504

13 Schwartz MF, Kimberg DY, Walker GM , et al. Anterior temporal involvement in semantic word retrieval: voxel-based lesion-symptom mapping evidence from aphasia. Brain 2009; 132 (Pt 12): 3411-3427

14 Schwartz MF, Faseyitan O, Kim J, Coslett HB. The dorsal stream contribution to phonological retrieval in object naming. Brain 2012; 135 (Pt 12): 3799-3814

15 Dell GS, Schwartz MF, Nozari N, Faseyitan O, Branch Coslett H. Voxel-based lesion-parameter mapping: identifying the neural correlates of a computational model of word production. Cognition 2013; 128 (3) 380-396

16 Macwhinney B, Fromm D, Forbes M, Holland A. AphasiaBank: methods for studying discourse. Aphasiology 2011; 25 (11) 1286-1307

17 Forbes MM, Fromm D, Macwhinney B. AphasiaBank: a resource for clinicians. Semin Speech Lang 2012; 33 (3) 217-222

18 MacWhinney B. The TalkBank Project. In Beal JC, Corrigan KP, Moisl HL, Eds. Creating and Digitizing Language Corpora: Synchronic Databases, Vol. 1. Houndmills, Basingstoke, England: Palgrave-Macmillan; 2007: 163-180

19 MacWhinney B. The CHILDES project: The database, Psychology Press. 2000

20 Granger S, Gilquin G, Meunier F , Eds. The Cambridge Handbook of Learner Corpus Research. Cambridge, UK: Cambridge University Press; 2015

21 Collaboration of Aphasia Trialists (CATs) . Available at: http://www.aphasiatrials.org/ . Accessed December 12, 2015

22 American Speech-Language Hearing Association . National Outcomes Measurement System: Adults in Healthcare–Acute Hospital National Data Report. Rockville, MD: 2011

23 Center for Rehabilitation Research for Large Databases (CRRLD) . Available at: http://rehabsciences.utmb.edu/cldr/ . Accessed on December 12, 2015

24 Holland AL, Weinberg P, Dittelman J. How to use apps clinically in the treatment of aphasia. Semin Speech Lang 2012; 33 (3) 223-233

25 Constant Therapy . Available at: https://constanttherapy.com/ . Accessed January 8, 2016

26 Talkpath, Lingraphica . Available at: http://www.aphasia.com/ . Accessed January 12, 2016

27 Morrison GE, Simone CM, Ng NF, Hardy JL. Reliability and validity of the NeuroCognitive Performance Test, a web-based neuropsychological assessment. Front Psychol 2015; 6: 1652

28 Basso A. Prognostic factors in aphasia. Aphasiology 1992; 6: 337-348

29 Morrison GE, Simone CM, Ng NF, Hardy JL. Reliability and validity of the NeuroCognitive Performance Test, a web-based neuropsychological assessment. Front Psychol 2000; 6