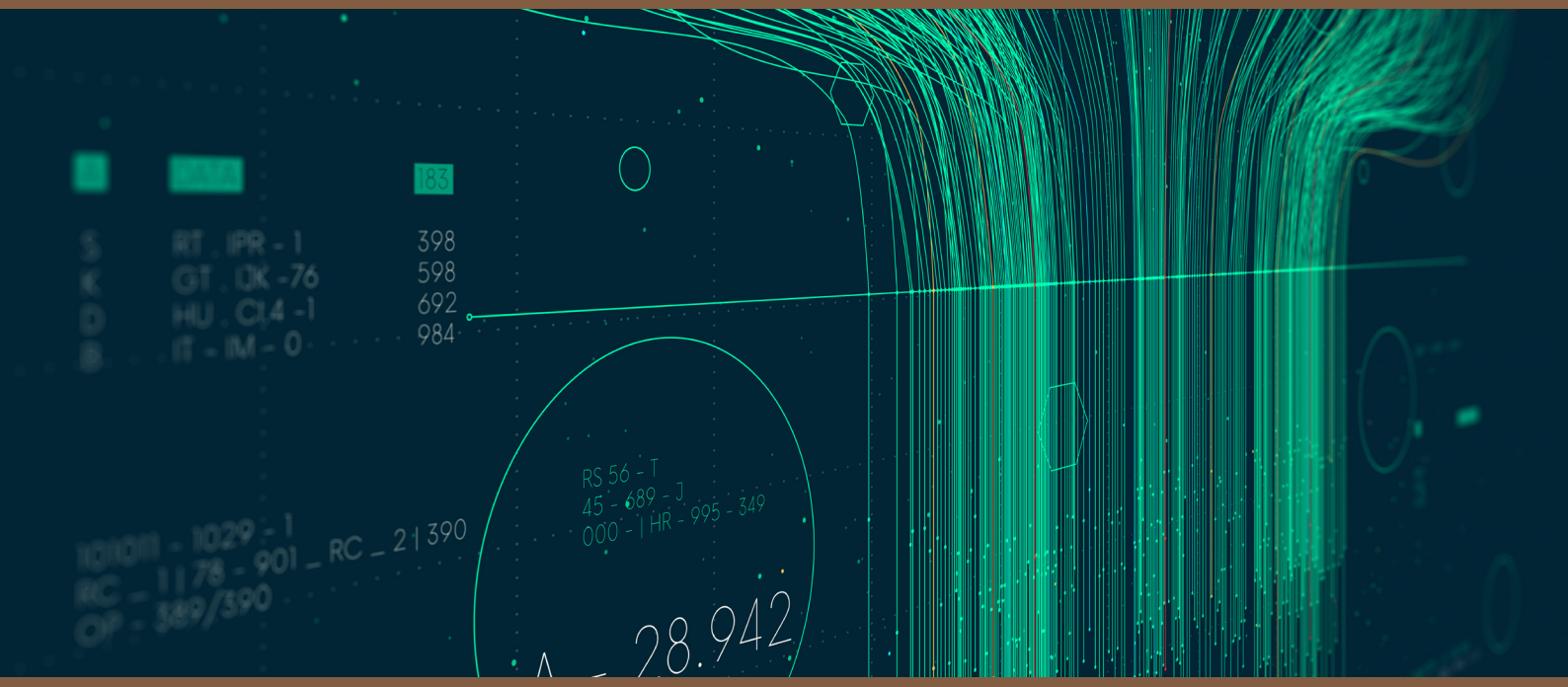
World Dementia Council
Leading the Global Action Against Dementia

# Global dialogue on data sharing for dementia research: Reflections

## The dementia landscape project

**Essays from international leaders in dementia**



Organized in partnership with

SageBionetworks

ADDI
Alzheimer's Disease Data Initiative

# Contents

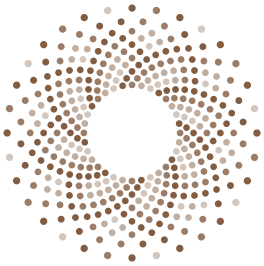
# 1. Introductions

## A decade of progress in data sharing for dementia research

**Dr Lara Mangravite**
President, Sage Bionetworks

Alzheimer's disease and related dementias are a major health burden across the world. These debilitating diseases have a devastating impact on patients, their caregivers and loved ones as well as on our healthcare systems. Despite the concerted efforts of researchers across the globe, our ability to advance strategies for dementia prevention, treatment, and monitoring are hindered by continued limitations in scientific understanding. As an example, we see increasing evidence that dementias are heterogeneous in nature - meaning that they can arise from multiple biological insults, present with multiple inter-related sets of pathologies, and progress with multiple clinical trajectories. Understanding these heterogeneitites is essential to address the individual needs of each patient. Large amounts of data are required to address this and many other unanswered questions in the field - more data than can be collected from any one source. Data-sharing is essential to meet these critical needs.

Over the past decade, the field has put enormous effort into the generation and coalition of data. In particular, we've seen an increased focus on data derived from humans. The seminal work of the Alzheimer's Disease Neuroimaging (ADNI) consortium, started in 2004, paved the way for open and rapid sharing of critical research data in dementia. Global initiatives continue to advance data sharing. For genomic studies, this is examplified by the International Genomics of Alzheimer's Project (IGAP). In addition, multiple epidemiological studies are coming together through initiatives including the Health Data Research UK, GAAIN, and the ADDI. Human molecular data - and the samples themselves - are being shared through a variety of programs including the NIA-funded NICRAD biorepository and the AMP-AD target discovery program. The availability of these data resources - and others in development - and the commitment to releasing data under FAIR principles provides a tremendous opportunity to reach research goals across the field. And yet, there is more to be done. This includes the sharing of existing data that is not yet available, the development and sharing of additional data to address unmet needs, and the integration of data across existing data repositories.



“How” we share data is an active topic of debate. Although fully open data sharing - where data is made accessible to all people for all uses in any compute environment - enables the widest range of research, it does so at the expense of data security/ privacy and research transparency. Fully open data sharing is also often not possible, particularly for human-derived data. A range of solutions have been developed to address data sharing needs - examples include cloud-based data enclaves, federated data stores, generation of synthetic data sets and containerized “model to data” sharing programs. These solutions vary in terms of who can access data, the breadth of analyses that can be supported, the degree of autonomy that independent users have for research ideas, and - importantly - in the financial and personnel costs of data sharing. This breadth of options has developed because there is not a singular data sharing solution that universally meets all research needs. The field has taken an “as open as possible” approach by developing a variety of data sharing initiatives each designed to meet a specific set of goals. Sharing methods are adapted accordingly. Interoperability across some of these initiatives will be needed - and complicated.

The most important component to consider in data sharing is the data itself. Although we hope that more data will answer many of our open questions, we know that data sharing is not going to solve all of our research needs. Data sharing programs must objectively evaluate the value of the data that they share in terms of data quality and in terms of data fitness for purpose. Understanding what open questions can be answered by combining existing data sources - and what questions will require new data types and/or sources - will help us to most effectively and efficiently mobilize resources to obtain the data required to address the pressing needs in dementia research.

# The good that can come from widespread data sharing

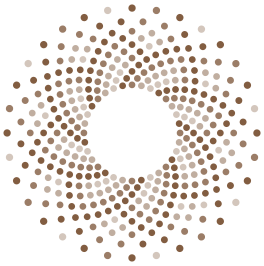
**Dr Tetsu Maruyama**
Executive Director, Alzheimer's Disease Data Initiative (ADDI)

Data sharing, as a concept, is exceptionally uncontroversial. Everyone agrees that having more eyes (and more computers) on data is a good thing and has the potential to lead to novel and important findings. When it is time to share data, however, things become more complicated. Many concerns surrounding sharing data (e.g., the risk of data being misused, the effort required to prepare the data for effective sharing) seem to outweigh the perceived benefits. It may, therefore, be worth outlining some of the good that can come from widespread data sharing, especially data related to understanding dementia.

Like all truly difficult problems, finding a cure for dementias will require innovative solutions. The best way to find these is to bring diverse thinking to the problem. There are many dimensions to diversity: different training and specialties (basic scientists, clinicians, data scientists, tool developers), diverse cultural, gender and language backgrounds, people with dementia and their caretakers and so on. Sharing data is a way to bring diverse people and skills together. An appropriately designed platform can allow people to participate without formal training, academic or industry affiliation, and without the level of funding that can often be a barrier to doing other types of research. The geophysicist in Argentina, the retired engineer in South Africa, and the undergraduate student in India can now all participate in generating and testing new ideas and sharing the results with a growing community. Sharing data creates communities.

An indirect consequence of increased sharing of dementia data will be a corresponding rise in the number of scientists from other fields entering the dementia research community. We are often reminded of the shortage of researchers working in dementia, relative to cancer or heart disease for example. Capacity building in neuroscience can be challenging and time consuming even given the recent and highly welcomed increases in funding for Alzheimer's and other dementias. But bringing experienced or early career scientists with expertise in other relevant fields such as cell biology or immunology into the dementia field may be easier when we give them the ability to work with large, integrated, well curated datasets. By enabling them to generate and even test hypotheses at low cost and with a low barrier to entry, these scientists may begin to see how they can contribute to curing dementia and move more of their efforts towards that goal.

As vital as the community will be to solving problems, data sharing is ultimately about data. Just as it is optimal for researchers to form communities that work together seamlessly to solve difficult problems, the data in turn should be fully interoperable.



The ability to combine data from different and disparate datasets into a single analysis should allow researchers to ask questions with sufficient scale and diversity to address key concerns such has heterogeneity in disease presentation and progression. This will allow scientists to generate and test hypotheses about diverse mechanisms for the initiation, development, and progression of dementias. But making datasets interoperable is far easier said than done. A commitment from research subjects, data contributors and data tool developers to work together towards a common goal will be essential. Those who become part of that community will be richly rewarded with new understandings and insights to fuel progress in eradicating dementias.

# 2. Challenges and opportunities

## The data challenge in aging and dementia research, clinical everyday life, and care: A search for solutions

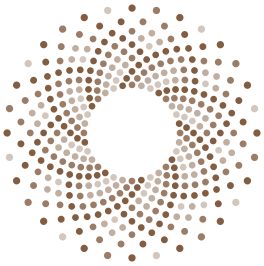
**Professor Joachim Schultz**
Director, Systems Medicine, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE)

**Data sciences will become much more critical in dementia research**

With the enormous increase in life expectancy at the same time leading to 'aging societies', we see a simultaneous increase of chronic disabling diseases particularly dementias in the elderly people. In contrast to diseases such as cancer or diabetes, for which molecular mechanisms are much easier to be studied and subsequently utilized to develop diagnostic biomarkers and therapeutic targets, diseases with the involvement of the brain are much more difficult to tackle. Finding solutions for patients with dementias requires a much more holistic undertaking including the acquisition of data ranging from the continuous observation of behavioral changes, movement sequences, high resolution images, to the measurement of surrogate tissues such as the blood or cerebrospinal fluid utilizing high throughput omics technologies with the highest resolution, down to the single cell level. In addition, current attempts to understand dementias on the molecular level, mainly in animal models need to be integrated into the holistic view.

**Some of the challenges in data sciences in medicine**

If we want to successfully utilize data generated from such different fields to improve the life of patients with dementia, we need to be able to utilize all this data, which to a large extent is generated in a very decentralized fashion. For example, health records might be available at the private practitioner's office, general diagnostic data at regional hospitals and more specialized diagnostics at specialized centers. These centers across regions, nations or worldwide are not connected in a seamless way. Currently, data exchange is only possible and organized in context of clinical trials or research studies with often cumbersome data exchange procedures.



The challenge is not only the decentralized fashion of data generation, but also the organization of such data and the legal requirements to which this data is subject. For example, data collected from patients in Europe require the recognition of the General Data Protection Regulation (GDPR), which are valid worldwide and not only within the member states of the European Union. At the same time, science requests 'Open Science', 'Open Data', 'Open Access' policies, which are important goals, while at the same time extremely challenging considering the current organizational state of most of the data and the legal requirements that have to be fulfilled to reach such high standards.

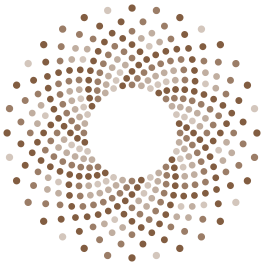
Beyond that, however, the question must be allowed as to whether complete 'Open Data' and completely 'Open Access' to all data is really the right way to go. Isn't it rather true that it's not about data per se, but about the insights from the data? Data without insights do have little value and data used in a wrong way by the wrong people can result in significant harm. Further, 'Open Data' and 'Open Access' contradicts very old medical traditions, namely the patient physician privilege. So, are we misguided at the moment concerning data sciences in medicine, because we follow the technological developments in the IT sector, rather than asking what we really need for medicine to benefit from data and information sciences?

**Principles for solutions**

Open Science is a must, but wouldn't it be better to exchange 'Open Data' with 'Open Insights'? This would mean that we do not share data freely but use the data jointly to gain novel insights from the data. How would one envision to do so? In this case, everybody who has valuable data pointing towards new findings about the pathophysiology of dementia, or disease progression, potential diagnostic markers or novel targets for therapy, would join into a network of scientists that would like to provide their data for joined learning without really sharing the own data. If the own data are not shared with others, but only the insights, data privacy laws are much easier to be dealt with. In addition, if patients would deny usage of their data for scientific purposes later, it would be so much easier to fulfill such requests, if the data would have never been duplicated and moved to another place.

**Simple solutions, partner institutions share insights instead of data**

The simplest model to do so, is when knowledge or insights from data has been gained at one institution and then a partnering institution would perform a validation analysis on their own data following the procedures and algorithms and analyses used at the initiating partner. The algorithms and procedures would follow the data and not the other way around. We have successfully illustrated this during the COVID-19 pandemic. This simple model of collaboration without the need to share data, but rather insights is even possible nowadays with single cell data in a clinical setting (Bernardes et al., 2020; Krämer et al., 2021; Schulte-Schrepping et al., 2020). We actually have applied this principle to several additional studies that are currently evaluated at prominent journals (Georg et al., 2021; Wendisch et al., 2021). Therefore, we are now setting up a framework at the DZNE to follow these principles in dementia research, when applying high-resolution single cell analysis to patient cohorts.



**Sophisticated solutions, the Swarm Learning principle**

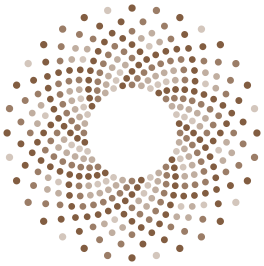
But we can already go beyond the simple model of sharing insights rather than sharing primary data. Imagine, you would have a computer infrastructure that holds your primary data on your premise at a secure place; further imagine you have a secure way of negotiating with partner institutions, how you would actually analyze the local data, how the data would have to be stored locally and made accessible only locally to locally run algorithms, how the results from local analyses (the insights) would be shared among network partners, how credit and responsibilities would be shared among network partners, and how you would report together on shared insights that come from such joined endeavors. Further imagine, how much more powerful we could utilize decentralized data, if we could make use of them as envisioned here. From computer sciences it is very clear that more data for machine learning, particularly when applying algorithms of artificial intelligence (AI), leads to better results. Imagine, how much better our data analysis could become, if we would join on learning together on our own data, while completely preserving data confidentiality and privacy locally. We call this vision the Swarm Learning principle. Equal partners with equal rights and responsibilities share insights derived from data that are private. Like the members of a swarm, here every member follows rules that apply to all swarm members simultaneously and equally.

**The pilot of the Swarm Learning principle: It is technically feasible**

Is this a vision only, or can we move forward and bring this vision to reality? Together with our partner Hewlett Packard Enterprise (HPE), the DZNE used the time during the pandemic and showed that the Swarm Learning principle is not only technically feasible, but it can be applied to medical data (Warnat-Herresthal et al., 2021). Swarm Learning is a decentralized machine-learning approach that unites edge computing, blockchain-based peer-to-peer networking and coordination while maintaining confidentiality without the need for a central coordinator, thereby going beyond federated learning. Using chest X-rays and blood transcriptomes, we illustrated the power of the Swarm Learning principle in applying AI to medical data in a diagnostic setting for leukemias, tuberculosis, COVID-19 as well as a number of pathophysiological findings in chest X-rays. We are now planning to extend the system to immune data as well as data in dementia research gathered at the DZNE and together with partner institutions. For us, Swarm Learning is not only a unique and exciting technical solution. It is a conceptual change since the IT solution follows the need of the medical field and not the other way around. Swarm Learning recognized old medical tradition, learning from each other, and keeping patient physician privilege intact. For us, it is therefore not only a new technical solution to make use of decentralized data, it is a changing point of how the IT sector should develop solutions that recognize and honor medical traditions.

**Summary**

Collectively, the chances to understand complex disease such as the dementias by utilizing every growing data from many different domains has never been better. The nature of decentralized data generation and the requirements for data privacy and protection requires to think computer infrastructures and approaches new. We think that Swarm Learning Principles are a major driver for this necessary change. At the



same time, we strongly advocate to change from the current concept of 'Open Science', 'Open Data', 'Open Access' to 'Open Science', 'Open Insights', 'Open Access to Insights' but strongly protect Data Privacy. In other words, 'Open Data' is not part of the concept forward anymore.

***Bibliography***


*Bernardes, J.P., Mishra, N., Tran, F., Bahmer, T., Best, L., Blase, J.I., Bordoni, D., Franzenburg, J., Geisen, U., Josephs-Spaulding, J., et al. (2020). Longitudinal Multi-omics Analyses Identify Responses of Megakaryocytes, Erythroid Cells, and Plasmablasts as Hallmarks of Severe COVID-19. Immunity 53, 1296–1314.e9.*

*Georg, P., Astaburuaga-García, R., Bonaguro, L., Brumhard, S., Michalick, L., Lippert, L.J., Kostevc, T., Gäbel, C., Schneider, M., Streitz, M., et al. (2021). Complement activation induces excessive T cell cytotoxicity in severe COVID-19. medRxiv.*

*Krämer, B., Knoll, R., Bonaguro, L., ToVinh, M., Raabe, J., Astaburuaga-García, R., Schulte-Schrepping, J., Kaiser, K.M., Rieke, G.J., Bischoff, J., et al. (2021). Persistent natural killer cell dysfunction in severe COVID-19. Immunity. 2021 Sep 4:S1074-7613(21)00365-4. doi: 10.1016/j.immuni.2021.09.002.*

*Schulte-Schrepping, J., Reusch, N., Paclik, D., Baßler, K., Schlickeiser, S., Zhang, B., Krämer, B., Krammer, T., Brumhard, S., Bonaguro, L., et al. (2020). Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. Cell 182, 1419–1440.e23.*

*Warnat-Herresthal, S., Schultze, H., Shastry, K.L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Händler, K., Pickkers, P., Aziz, N.A., et al. (2021). Swarm Learning for decentralized and confidential clinical machine learning. Nature 594, 265–270.*

*Wendisch, D., Dietrich, O., Mari, T., von Stillfried, S., Ibarra, I.L., Mittermaier, M., Mache, C., Chua, R.L., Knoll, R., Timm, S., et al. (2021). SARS-CoV-2 infection triggers profibrotic macrophage responses and lung fibrosis . Cell, 2021, accepted.*

# A global knowledge graph for integrated brain health research data

**Dr Sean Hill**
Director, Krembil Centre for Neuroinformatics,
Centre for Addiction and Mental Health, Toronto, Canada

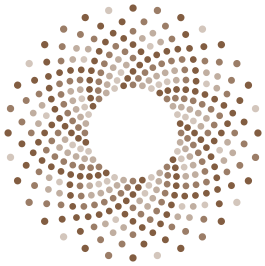
Billions of dollars are spent on brain health research around the world every year. However, the data produced from this research is often not available, discoverable or reusable, severely diminishing their impact. Researchers are unable to rapidly find relevant, high quality data that can be reused for new analyses. Further, the data are often stored in isolated siloes, and highly heterogeneous, making a comprehensive inventory of available data and the key gaps to be addressed virtually impossible. The FAIR principles – making data Findable, Accessible, Interoperable, and Reusable – outline the requirements for making it possible to integrate data from around the world. However, the technical, sociological and incentive framework for readily discovering, reusing and integrating brain health data remains a major challenge.

The World Wide Web, originally developed in the early 1990s at the CERN to enable information-sharing between scientists around the globe, has transformed the world by providing the underlying standards and protocols for hyperlinked multimedia content. By 1998, Google was established with a mission to “organize the world's information and make it universally accessible and useful”.

Today, Google uses the standards of the World Wide Web and the data schemas of schema.org (founded by Google, Yahoo, Microsoft and Yandex) to create an immense knowledge graph that represents organized data and knowledge constructed from sites across the web. A knowledge graph is a type of database that contains data about real world entities (people, places, things, or concepts) and puts this data in context through interlinked relationships to other data. Because a knowledge graph contains machine-readable representations of data and their context it can enable automated reasoning, inference and serve as the basis of powerful artificial intelligence applications.

The Google Knowledge Graph provides the basis of the powerful search engine and AI capabilities that enable up-to-the minute discovery of people, places, products, and businesses across every industry around the globe. Google employs a hidden business model where the user pays nothing, while the company makes billions in revenue selling advertisements. Today, if information about a person, place, product or business cannot be discovered through Google, they effectively do not exist for most of the world.

Today, the primary mechanism for a researcher in the biomedical or life sciences to discover the latest scientific results is to search PubMed.gov, which enables keyword searches of the vast array of publications across these domains. Scientific papers contain

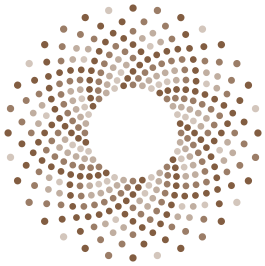


(usually) human readable descriptions of scientific results with the full diversity of linguistic expression. This usually means that papers need to be read to fully appreciate the relevance and significance of a scientific finding and situating it in context requires significant expertise in each domain. Scientific career progression is dependent on impact factors and publications. As the number of publications have exploded in recent years, it has become increasingly difficult to stay up-to-date with the latest scientific knowledge by reading all relevant publications. Furthermore, the ability to form an integrated picture of our current knowledge across neuroscience is now well beyond the scope of any individual, making the most fruitful path for novel discoveries and career success a narrow focus on a specialized field.

Discovering and accessing the data that underpins new discoveries also remains a significant obstacle to rapid scientific progress. Today, the reward for a new publication is far greater than the effort required for effective data sharing and reuse. Although, data sharing is increasingly required by life science and biomedical funders it requires significant resources to ensure data are sufficiently documented and structured for reuse, which is rarely funded. With limited resources, scientists are best served to focus on proprietary, novel discoveries that are often underpowered. In addition, high impact journals prioritize novel findings with new techniques which disincentives generating data that can be readily combined with other studies. The net result of the current incentive system is a large number of individual studies, with insufficient statistical power, containing data that cannot readily be aligned or integrated with other studies.

Scientific data reuse requires a deep understanding of the process that generated the data. Methods sections in papers are rarely (if ever) sufficient to understand and reuse a dataset. Data structures and descriptions for dissemination rarely reflect the study process, making it challenging to truly understand the meaning or significance of a dataset. Scientists who share data often prefer to do so in the context of a collaboration so that the data are not misused or misinterpreted to generate misleading or false findings. In addition, a collaboration typically ensures a co-authorship of a publication thus providing value for the time invested.

Provenance information, metadata that describes the context and process of how the data was produced, using which technique, and by whom (or what), is essential to enable the automated reuse and integration of data. Neuroshapes.org provides examples of neuroscience data schemas (modeled on existing schema.org and W3C provenance standards) that include linked representations of the provenance of diverse data types including electrophysiological recordings, neuron morphology reconstructions, and brain atlases. These representations capture differences between different data types based on their provenance. For example, not all neuron morphology data are identical. Some neuron morphologies are generated from brain slices of a specific thickness, where the cells are accessed intracellularly using whole cell patch clamp techniques, injected with a dye, then histologically prepared for optical imaging and manual reconstruction. Other neuron morphologies are generated from a genetically engineered mouse line, labeled with viral injections, and imaged using fluorescent light sheet microscopy before being automatically reconstructed by a machine learning algorithm. The validity of extracted features of the neuron morphologies (e,g. spine count, sizes, dendrite diameters, etc.) will depend on the details of the specific processes. Thus, this information is critical to inform whether these data can be legitimately used to answer a specific research question.

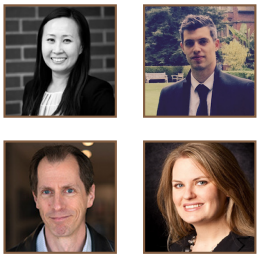


Provenance information can also be used to track data reuse and give credit to all the individuals and organizations involved in generating the scientific discovery. For example, provenance information can contain the full lineage of the data as described above, giving the detailed processes that produced the data, but also giving credit to all the individuals (trainees, scientists, technicians) and organizations (labs, institutions, companies, and funders) involved. In addition, when novel data processing algorithms are involved, ensuring that those who created the algorithms are credited for their impact can help sustain the development of these important tools. Published, machine-readable, provenance information can therefore be a powerful tool in establishing a more comprehensive and inclusive incentive system.

Sensitive data requires careful data governance that ensures the data are used only according to the granted ethical approvals. A revocable drivers licence for data (the GA4GH has recently established an impressive framework for data passports with visas) that is granted to accredited researchers with appropriate training and certification for handling sensitive data can help streamline access to these data. However, the issues of discovery of data and access can be made separable – such that knowledge of all available data are public, but the access to the sensitive data are controlled and regulated to approved researchers. Machine-readable information about the governance of the data including ethics status, license and data use agreements can make it much easier to identify appropriate and accessible data for research.

A global knowledge graph that integrates brain biomedical research data and includes rich provenance information would dramatically transform and accelerate brain health research. It could serve as the basis of a rich ecosystem of tools, much as we have seen in the tools built on the Google Knowledge Graph: atlases, search engine, marketplace, AI tools, and more. Imagine a global cohort explorer to find, access and analyze integrated study data from around the world. Notification systems could inform researchers when new data similar to their own data has been published. Researchers that analyzed existing datasets could be notified when new data that could add statistical power to their study have become available. Machine learning tools could query and mine the data for new predictions. Funders could hold challenges with large datasets (much as Kaggle and Sage Bionetworks have pioneered), impact leaderboards could highlight the latest contributions from the global community. Ultimately, the knowledge graph could help identify key gaps in available data and knowledge and help funders prepare targeted funding to address these gaps.

While the technical standards to establish such a knowledge graph exist, the resources, leadership, and governance are not yet in place. The transformative power of organizing data and making it universally accessible and useful is well established by the example of the World Wide Web and Google. The technical standards for publishing standardized well-structured metadata (following the existing models of schema.org, bioschemas.org, and neuroshapes.org) to a public web site are also well established and sufficient to enable a web crawler to construct a knowledge graph. However, it will take a well funded and focused effort to establish the metadata standards and their governance for the life sciences and biomedical domains, the tooling to easily capture, edit, and publish them, and the infrastructure to operate the knowledge graph and search engine. Most of all, such an effort requires strong incentives to ensure researcher adoption and the creation of an ecosystem for integrating our global data and knowledge.



# Data sharing with the Global Alzheimer's Association Interactive Network

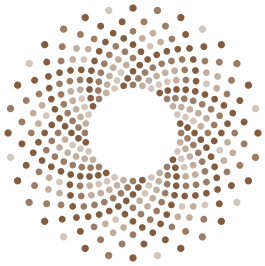
**Dr Cally Xiao,**
**Dr Ioannis Pappas,**
**Dr Scott C. Neu,**
**Dr Heather Snyder,**
**and Dr Arthur W. Toga**
The Global Alzheimer's Association Interactive Network

Recent technological advances in experimental techniques have contributed significant amounts of data to study Alzheimer's disease (AD) and related dementias (ADRD). However, these advances made by individual groups from isolated institutions present only a fraction of the story. Data sharing allows for expanding the potential of the collected data by increasing cohort size and identifying new trends from existing studies. Data sharing also promotes transparency and reproducibility in research. Thus, it is of utmost importance to unite separate research efforts into a central hub to manage data with the goal of promoting data sharing and collaboration to study complex diseases such as AD/ADRD.

There are, however, challenges in data sharing. First, the privacy of research participants in the studies must be protected by not including individual identifying information during the data sharing process. Second, a data sharing platform should be able to withstand the complexity of data collected, such as imaging, biomarkers, proteomics, and genetics data while harmonizing naming conventions across different studies. Finally, data owners may be sensitive to control over data and feel that control should remain with the researchers who collected the data and their institutes.

The Global Alzheimer's Association Interactive Network (GAAIN, www.gaain.org), established in 2015, is the world's first federated network connecting independently operated AD/ADRD data repositories from around the world. GAAIN addresses the need for a global virtual community to share AD/ADRD data. GAAIN is designed to overcome the challenges in data sharing and to evolve with the technological advances in data collection. As a resource for researchers to discover and explore studies to use in secondary analyses, GAAIN has connected over 50 studies from 17 different countries in a network consisting of over 500,000 subjects and is currently the only AD data sharing platform that allows researchers to preview existing data and perform analyses without prior programming knowledge.

The GAAIN platform offers flexibility to accommodate new and updated data sets while respecting data ownership and privacy concerns. First, any information that could potentially be used to identify subjects are removed by the researchers. In addition, the GAAIN platform can process any type of data that can fit in a spreadsheet, and currently the categories of variables shared in GAAIN are demographic, clinical, cognitive,



genetic, imaging-derived, environmental or behavioral information, laboratory results, and family and medical history. GAAIN also harmonizes nomenclature from different studies that have common attributes such as diagnosis and APOE genotype. Finally, studies in GAAIN still belong to the data owners and they would review data requests from GAAIN users who are interested in full data sets.

GAAIN users can access, visualize, and analyze the linked data sets via a client-server network while the data resides in the data owners' institutes. The flagship tool of the platform is the GAAIN Interrogator, which allows for cohort discovery, data set aggregation, data visualization, and preliminary analyses. As an incentive to further protect ownership, the analysis methods used by the Interrogator and the underlying analytical models are not made publicly available, thus researchers would not be able to directly publish the results from the Interrogator. Researchers using GAAIN can only obtain the full data sets via direct requests to the data owners before performing formal analyses for publications.

In the GAAIN Interrogator, researchers can analyze both individual and pooled data sets. Further, they can define their own variables and create custom cohorts for further analyses. Analyses include odds ratios from logistic regressions, correlations from linear regressions, survival fractions from Cox regressions, and in the case of pooled datasets, weighted odds ratios from Mantel-Haenszel meta-analyses.

Researchers who have used GAAIN for their meta-analysis studies were able to leverage and combine existing data to answer questions that would not be possible when analyzing one data set alone. For example, in a meta-analysis study using data sets discovered in GAAIN, researchers found that women with the APOE ϵ3/ϵ4 genotype had an increased risk of developing AD compared to men between 65 and 75 years of age but had similar risks overall between 55 and 85 years of age (Neu et al., 2017). In a study presented at the Alzheimer's Association International Conference (AAIC) in 2020, researchers used two data sets from GAAIN to find that, surprisingly, alcohol use was associated with better cognitive performance (Funk-White et al., AAIC20). In another study presented at AAIC 2020, a researcher examined MRI scans of subjects who are diagnosed with mild cognitive impairment (MCI) and have suspected non-amyloid pathology (SNAP-MCI), combined from four data sets in GAAIN. This researcher discovered that SNAP-MCI subjects had faster atrophy rates in the frontal cortex than MCI subjects, although there were no differences in cortical thickness between the two groups after correcting for site-related variations (Rane, AAIC20). At the recent AAIC 2021, the GAAIN team presented findings that Latino/Hispanic individuals with the APOEϵ4 allele were less likely to develop MCI or AD compared to non-Hispanic white individuals with the APOEϵ4 allele (Xiao et al., AAIC21) in a larger study sample than previous reports.

GAAIN provides data analysis tools, enables pooling of data sets, and has brought together data from longitudinal studies, cross-sectional studies, and clinical trials. The leadership of GAAIN works directly in AD/ADRD research. With a deep understanding of the field and research community, the GAAIN team is constantly improving the GAAIN platform based on new technological developments, data needs of the community, and feedback from researchers and data owners. Our vision for the future of GAAIN involves more widespread usage and awareness, where researchers interested in a

particular question related to AD/ADRD first look to GAAIN to explore trends in linked data sets and then design meta-analysis studies to support or rebut previous findings and visualize study trends on a larger scale. Global data sharing would allow for deeper insights into study trends across countries and cultures, better understanding of risk factors of AD/ADRD, and increased collaborations amongst researchers who are working towards effective treatments and preventative options for those suffering from or at risk of AD/ADRD.

# Data sharing in Alzheimer's disease: the EMIF-AD experience

**Dr Pieter Jelle Visser**
Clinical Epidemiologist, Maastricht University Medical Centre and VU University Medical Center Amsterdam

**1. Why data sharing?**

Data sharing is an efficient way to increase sample sizes which will improve statistical power. Data sharing will also enable cross-validation of results in independent cohorts. It may also help to estimate disease prevalence on a population level rather than the cohort level. Moreover, data sharing maximizes the use of data, as in most cohort studies, local researchers can only analyze a fraction of the data collected.

**2. Type of data**

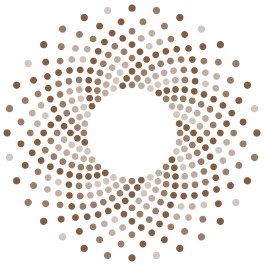
There are three main types of healthcare data: 1. data collected as part of dedicated research cohorts; 2. data collected as part of routine care; and 3. a combination of both. The type of data determines how data can be shared. For example, research cohort data are collected with consent of the research participants, and depending on the consent given, subject-level data can be shared with other research groups. No consent is typically available for sharing of data collected as part of routine care. This type of data typically needs to be analysed remotely in secure data environments, which limits the possibilities to pool data from different sources. The advantage of this type of data is the large sample size and sometimes long follow-up. A limitation is that routine data are often not collected in a systematic way and lack detailed phenotyping, for example by scans or body fluid markers.

**3. How to share data?**

Data sharing typically involves several steps: identification of cohorts that could provide the data needed. This can be done with meta-data catalogues, such as the EMIF-AD catalogue. The next step is to set-up a research collaboration and data or sample transfer agreement. Then, if data need to be pooled across cohorts, the data from different cohorts needs to be reformatted in common data format.

**4. Data sharing strategies**

One could have different approaches to share research cohort data. The first one is to regularly upload cohort data to a central platform on which data can be pooled and



analysed. The other approach is to keep data locally, and only upload or pool data if needed. Data generated as part of a common research project, could then be shared through a central platform. This latter approach was applied in EMIF-AD.

**5. The EMIF-AD experience**

The EMIF project from the Innovative Medicines Initiative (IMI) aimed to facilitate reuse of data to address scientific challenges in common diseases, including Alzheimer's disease (EMIF-AD). In the EMIF-AD multimodality biomarker discovery study, we investigated biomarkers for diagnosis and prognosis of AD in the predementia stage (Bos et al. Alzheimer's Research & Therapy (2018) 10:64 https://doi.org/10.1186/s13195-018-0396-5). Through the EMIF catalogue, we identified 10 cohorts that could provide MRI scans, CSF, plasma and DNA samples and clinical data. It took around 30 months until all contracts were signed. We eventually included data from 1200 individuals and performed in subsets ranging from 400 to 1000 individuals, plasma proteomics and metabolomics, CSF proteomics, cortical thickness analysis, and GWAS and exome analysis. Data were pooled using the EMIF-AD ontology. The biomarker data and clinical data were uploaded on the tranSMART platform from which they can be shared.

In the EMIF-AD preclin-AD study we enriched data from an existing cohort study, the Netherlands Twin Registry, with AD biomarkers in CSF and MRI scans, and amyloid PET scans (Konijnenberg Alzheimers Res Ther. 2018, doi: 10.1186/s13195-018-0406-7).

In another EMIF-AD study, we pooled existing data from over 50 cohorts to estimate the prevalence of amyloid pathology in non-demented and demented individuals (Jansen et al JAMA. 2015, doi:10.1001/jama.2015.4668; Ossenkoppele et al JAMA. 2015, doi:10.1001/jama.2015.4669. Taken together, EMIF-AD showed the feasibility of pooling existing cohort data and samples in the field of AD (for other EMIF-AD examples see https://pubmed.ncbi.nlm.nih.gov/?term=emif-ad).

**6. Challenges**

With the GDPR regulations, data sharing has become more challenging form a legal and ethical perspective. Another challenge is to incentivize data sharing. While researchers are generally happy to share their data, it takes time to do so, and future data sharing projects should consider providing funds to cohorts to support the data sharing.

**7. Future perspectives**

The need of data sharing is now widely acknowledged, and several initiatives have been launched to support this recently. The Alzheimer's Disease Data Initiative has developed a platform for identification of data and central storage of research data. ADDI technology will also be used in a new IMI project on data sharing: the European Platform for Neurodegenerative Disorders (EPND), which started November 1st 2021. The aim of EPND is to combine existing data sharing infrastructures in Europe (including the EMIF catalogue for example) to enable biomarker discovery in Alzheimer's disease and Parkinson's disease using data from over 60 cohorts. One of the data analyses tools that will be implemented is federated access to local databases. This would allow remote analysis of data stored locally such that data do not to need to be leave the cohort database.

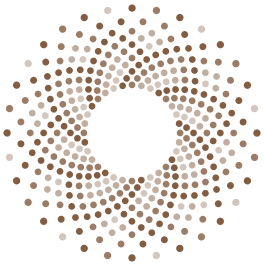
# Data sharing platforms: the gateway to dementia disease trial data

**Dr Rebecca Li**
Executive Director, Vivli, and on faculty at
the Center for Bioethics, Harvard Medical School

In recent years, the culture of clinical research has progressively gained momentum towards one of increased transparency and accountability. Individual participant-level clinical trial data (IPD) sharing is the latest of major efforts on trial transparency that started with clinical trial registration. IPD sharing of clinical trial data has now become a reality thanks to the significant nudge from journal editors, funders and others such as the NIH who have begun to initiate stronger policies supporting the re-use and sharing of individual participant-level clinical trial data (IPD). Together, these policies have led to a new era of opportunity for transparency. Researchers can no longer leave results in unpublished anonymity regardless of the outcome but must make public the clinical trial summary results. And increasingly, researchers are being asked to share a study's individual participant-level data. Access to this participant level or "raw" data from individual participants enables researchers to combine data to drive new scientific insights and assess reproducibility in ways simply not possible with summary or aggregate data tables. For dementia, this may mean that new trials in development could leverage key learnings from prior trials to inform the design of the protocol, generate new scientific hypotheses, integrate and combine data from multiple studies (IPD meta-analyses) to drive forward new insights or re-analyze trial data (to validate or confirm or reproduce existing findings).

Data could be accessed from trial platforms or repositories and the number of clinical trial data sharing platforms has grown and many remain disconnected from each other resulting in a fragmented ecosystem. Within the ecosystem of current data sharing repositories and platforms, there is a spectrum of approaches to execution. Some platforms and repositories employ an "open access" or "open data" approach allowing the data are to be available via download upon signing of a legal use agreement. Other platforms offer a "managed access" or "gatekeeper" model whereby additional requirements and elements are typically needed for data access including submission of a research proposal, signing of a data use agreement and review through an independent panel that typically reviews proposals for scientific merit.

The FAIR Data Principles are foundational and serve as guide for research data sharing. Under FAIR, data Data should be identifiable and searchable ("Findable"), accessible under ("Accessible"), able to be combined with other data in a meaningful way ("Interoperable"), and able to be re-used for multiple purposes ("Reusable"). Vivli, a global data sharing platform established 3 years ago exemplifies a practical implementation of these FAIR data principles and currently provides access to clinical



trial data for re-use from 38 institutional members Vivli.org including 23 from the biopharmaceutical industry and 15 academic institutions, government, non-profit organizations. Currently Vivli provides access to over 6,300 trials which represents 3.6 Million trial participants from over 120 countries. Included in the repository are numerous trials in Alzheimer's Disease and other dementias that may be integrated to drive "big data insights" through the combining of these multiple studies (or with data from other diseases to answer questions across diseases such as examining the interaction of Alzheimer's Disease and Covid-19, Alzheimer's and Type 2 Diabetes, or dementia and Parkinson's Disease) and pooling these analyses to answer specific questions.

The increasing prevalence of dementias such as Alzheimer's Disease has only magnified the necessity for these data sharing platforms and the need for data sharing. If most data are shared in a timely fashion that ideally allows for aggregation and analysis of datasets, this increases the probability of the community moving towards a more cohesive scientific understanding of the causal mechanisms of the dementia disease state. Conversely, if datasets are not shared or platforms exist as silos, society could lose an opportunity to develop key insights into our understanding of dementia and its associated challenges. However, for data sharing to be enthusiastically embraced by those doing the sharing, there needs to be tangible benefits to data contributors and the scientific ecosystem. These tangible benefits will serve to tip the culture in favor of data sharing. In the end it is important to remember that data sharing is not the goal in of itself however, but it is the outcomes that data sharing yield that ultimately may benefit individual dementia patients and the larger community as a whole.

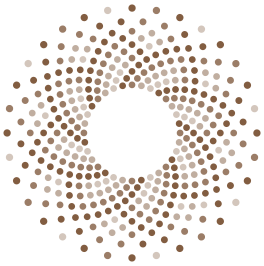
# The current academic culture: an intangible barrier to data sharing for a true collaborative effort against dementia


## Professor Yves Joanette

Professor of Cognitive Neurosciences and Aging, Faculty of Medicine, Université de Montréal


In 2012, the Prime Minister David Cameron took advantage of a G7 meeting in London to launch the Global Action Against Dementia, an initiative that would later evolve into the World Dementia Council. The motive was clear: given the rapidly increasing burden of dementia on people, health systems and countries' finance, there was an urgent need for collaboration between all those involved, from citizen to researchers and funding agencies, and from industry to governments and global organisations, such as WHO. The need for such a collaborative approach was judged urgent since it was recognized that, no country or sector alone would ever be able to crack the code of dementia not only because of the magnitude of the challenge and the required financial investments, but mostly because of the necessity to bring together intelligentsia from academia and industry. The goal of this collaborative effort was to unveil the causes of the neurodegenerative diseases causing dementia to propose disease-modifying treatments, as well as to better understand how to best support those living with dementia, their caregivers, as well as the health systems. Some years later, in 2015, Margaret Chan, then WHO Director general, re-affirmed the necessity to collaborate between countries, between industry and academia as well as between academics. The message for enhanced collaboration was -- and is still -- clear and acknowledged by all.

However, one major intangible barrier to full-blown scientific collaboration in academia remains the culture of data ownership, despite the recognized efforts of some institutions and consortia, some of them in my own country. The challenges of implementing a true data-sharing culture in academia has been identified and discussed over and over during the last decades. But despite the academic attention devoted to the necessity to engage in data sharing, the day-to-day relationship with data has not changed that much in most academic milieu. A typical academic would still today refer to the output of her/his research as being her/his own data, without referring to the fact that she/he is essentially the steward of those data, for which the institution or the research centre has to act as the fiduciary. Such an attitude is not surprising given the fact that, forever, academia as well as funding agencies have incented personal accomplishments and outputs over collaborative efforts and data sharing. Over the last decades, policies have been adopted to encourage data sharing. For example, the Canadian Institute of Health Research has a publication and data sharing policy in place since 2015, and the Chief Scientist of Québec has been promoting Open Science for many years now. But there is still a long way from a policy which involve a culture change.



Interoperability, technology, regulatory and ethical contexts are all well recognized barriers to global sharing of data for research on dementia or other health challenges. However, it could be argued that the culture of data ownership amongst academic may represents the biggest intangible challenge currently. To tackle this challenge and ensure a culture of true data sharing in academia will require several convergent actions including:

- To educate researchers, research professionals and trainees on the reality about data as a public common that they must steward on behalf of the fiduciary corresponding to the university or the research centre, re-emphasizing that intellectual property, where applicable, is on the innovative ideas emerging from their own data or those shared by colleagues.

- To invest in data service organizations and alliances to help defining preferences for non-proprietary international and community standards that would facilitate access and provide guidance and training for the use and the interpretation of data along their lifecycle, provide domain specific data infrastructure, and endorse trustworthy repositories.

- To ensure that academic governance introduces clear incentives regarding data and publication sharing not only in their policies, but mostly as part of the evaluation processes for promotion and the recognition of excellence.

- To invite both public and private funding organisations to accompany their publication and data sharing policies with specific expectations and penalties in cases of non-compliance with these policies.

- To inform the public about the benefits of data sharing as a crucial measure that would facilitate unveiling the mysteries of complex health challenges, such as dementia.

Achieving an evolution of the academic culture towards a true data sharing culture and environment will not be an easy task. But the academic world that would result from such a change would lay the ground for the collaborative efforts required to tackle one of humanity's most complex public health challenge.

*The author would like to thank Camille Tremblay, Principal coordinator of the Consortium Santé numérique, as well as Aubert Landry, Special Advisor to the Research and Teaching Digital Strategy both at Université de Montréal, for having contributed to the ideas expressed in this essay.*

# 3. Concluding thoughts

## The encouraging progress of diagnostics and data sharing in dementia

**Bill Gates**
Co-founder, Bill & Melinda Gates Foundation

We've all heard the statistics: Today, nearly 55 million people suffer from Alzheimer's disease or related dementias. According to the WHO, dementia is the fastest growing burden on healthcare systems and the seventh leading cause of death worldwide. Its estimated cost to society is more than US$1.3 trillion every year—and it could be twice that by 2030.

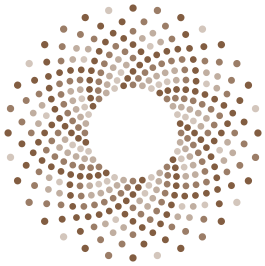
It can be difficult to comprehend the true toll these numbers represent, but it quickly comes into focus for anyone who has ever cared for or watched a loved one suffering from dementia. I lost my dad to Alzheimer's one year ago, so I understand firsthand the anguish caused by a disease for which there is still no effective treatment.

And yet I have never been more optimistic about the progress being made in Alzheimer's research, and the breakthroughs that may someday soon let us substantially alter the course of the disease. Finding a cure for Alzheimer's requires sustained innovation in five critical areas. We need to understand the underlying causes of the disease; improve detection and diagnosis; increase the pipeline of therapeutics; and remove barriers, especially for people from underrepresented populations, to enroll in clinical trials. Finally, to accelerate advances in all these areas, the global community must facilitate more and better access to data.

In the past few years, we have seen tremendous progress on all fronts. But I'm highly encouraged about two areas in particular: diagnostics and data sharing.

**Diagnostics**

Three and a half years ago, I partnered with the Alzheimer's Drug Discovery Foundation and a number of other investors to create the Diagnostics Accelerator. Since its inception, the Diagnostics Accelerator has committed $50 million in more than 35



projects focused on developing reliable, affordable biomarker tests for Alzheimer's, with promising results. For example, researchers at the University of Gothenburg are working on a method for measuring beta amyloid protein fragments in the blood. The medical imaging company RetiSpec has found a way to use hyperspectral imaging to detect beta amyloid in the retina. Cogstate, Altoida, and others are developing digital tools that can help monitor psychological and behavioral changes associated with neurodegenerative disorders.

I'm particularly excited about the GAP Foundation's recently announced Bio-Hermes study. It will allow researchers to evaluate new ideas for diagnostics—such as blood and ocular biomarker tests and cell phone apps—by comparing their results with the results of proven approaches, such as brain amyloid PET scans and traditional cognitive tests. This will accelerate the search for new diagnostics.

I'm optimistic that we'll soon have a number of inexpensive, non-invasive methods for diagnosing Alzheimer's and other dementias much earlier in their development.

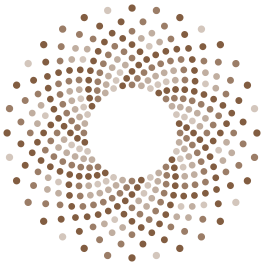
**Collaboration and data sharing**

It's encouraging to see the speed and scale of the research that's currently underway. To accelerate the pace of discovery, however, we must do more to maximize data sharing and collaboration. Sharing data allows researchers to look at the information from many studies at the same time, with the possibility of testing hypotheses or finding patterns and pathways that the original investigator wasn't looking for.

A few years ago, a coalition of organizations—industry, academia, advocates, and government—came together to address this need. The effort to facilitate more and better access to data culminated in the launch of the Alzheimer's Disease Data Initiative (ADDI). ADDI's focus is three-fold: increase the secure sharing of dementia-related data from academic and industry sources, make it easier to share data across platforms around the world, and empower researchers to find, combine, and analyze data that could lead to new discoveries.

Today, more than 2,000 researchers from 80 countries and regions have embraced the ability to work with multiple datasets in ADDI's secure environment, and that number is growing every day.

Data sharing also helps bridge the diversity gap that is inherent in most patient datasets. Alzheimer's is not limited to any one country, culture, or race. Yet, all too often, researchers are working with datasets from predominantly white patient groups, especially in North America, Western Europe, and Australia. If we can increase access to data from studies in Africa, East and South Asia, Latin and South America, or those focusing on a diverse group of participants, we can further our approach to this global disease. In the U.S., the Bio-Hermes study has committed to ensuring that 20% of its study participants are African American or Hispanic, setting the standard for future research and development.



**The path forward**

All of us can play a part in these efforts. If you're a researcher, thank you—I hope you'll continue your work on new diagnostics, treatments, and cures. If you are living with Alzheimer's or have a loved one who is, I hope you'll consider participating in a trial or study. As new diagnostics come along, it will be easier to know whether you're a good match for a trial. Finally, governments, industry groups, and foundations should keep funding research and development, and invest even more.

I believe that humans are capable of solving almost any problem through innovation—and I see countless opportunities for innovation in this field. I'm optimistic that if we all work together, we can defeat dementia.

The World Dementia Council (WDC) is an international charity. It consists of senior experts and leaders drawn from research, academia, industry, governments and NGOs in both high-income and low- and middle-income countries, including two leaders with a personal dementia diagnosis. The WDC has an executive team based in London, UK.

**worlddementiacouncil.org**


UK charity registration number: 1170743

Cover image editorial credit: Shutterstock.com