Paper DH11

Challenges and Wins of Pooling Internal and External Alzheimer's Data for Secondary Use

Michaela Mertes, F. Hoffmann-La Roche Ltd, Basel, Switzerland
Iska Gans, F. Hoffmann-La Roche Ltd, Basel, Switzerland

ABSTRACT

This presentation will give an overview of a pooled dataset including internal and external Alzheimer's study data for secondary use. Detailing the process to get all the data into a standardized and harmonized structure and content explaining some of the challenges. The final goal of this dataset is to set the foundation for elaborating and optimizing the analysis for scientists. So far 35 studies are included with over 75 000 subjects (including screen failures).

INTRODUCTION

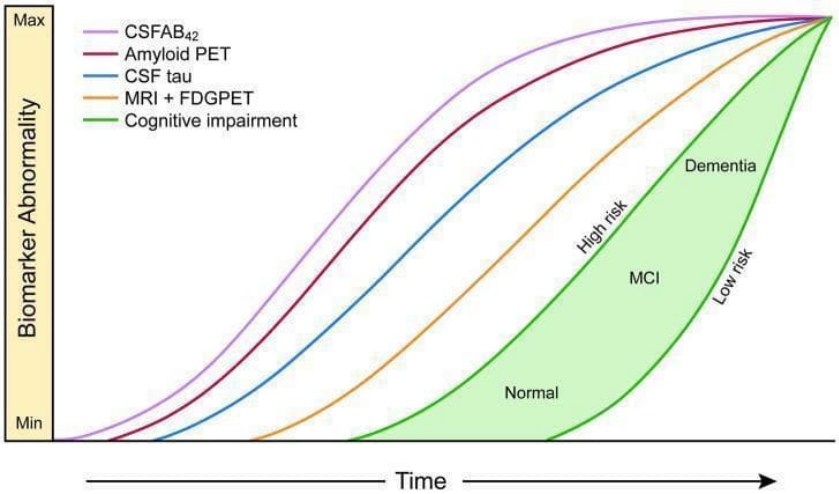
Alzheimer's disease remains a huge unmet need with major challenges in developing medicines and new diagnostics. With the FAIRification (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable) of in-house study data and external data assets, data driven decisions and learnings are improved. This paper provides a high-level overview of Alzheimer's disease and describes the creation of the data mart for secondary use including some challenges and wins.

ALZHEIMER'S DISEASE

Alzheimer's disease is a progressive neurological disorder that causes the brain to shrink (atrophy) and brain cells to die. It is the most common cause of dementia; accounting for 60-80%, estimated. Worldwide more than 55 million people live with Alzheimer's and other dementias. The disease is irreversible and slowly destroys memory and thinking skills. Behavioral and social skills decline, which affects a person's ability to function independently.

Alzheimer's disease typically begins to develop 20 years or more before symptoms become noticeable. It most commonly affects people aged 65 or older. The disease progresses slowly, and early signs can be subtle and easily mistaken by normal age-related changes.

The following figure illustrates that prior to cognitive impairment/ dementia biomarker values in cerebrospinal fluid (CSF) (CSF Abeta 1-42, CSF tau) change and position emission tomography (PET) scans show abnormalities.



Lancet Neurology vol. 12; Jack CR Jr, et.al, pp207-216;2013

ASSOCIATED BRAIN CHANGES OF ALZHEIMER'S DISEASE

Changes in the brain linked to Alzheimer's disease can be associated with the following pathological hallmarks:

1. AMYLOID PLAQUES

Clusters of protein fragments called beta-amyloid (amyloid plaques) accumulate in the brain outside neurons. Neurons connect with each other through synapses. These connections are vital for transmitting information. These plaques disrupt the communication between nerve cells.

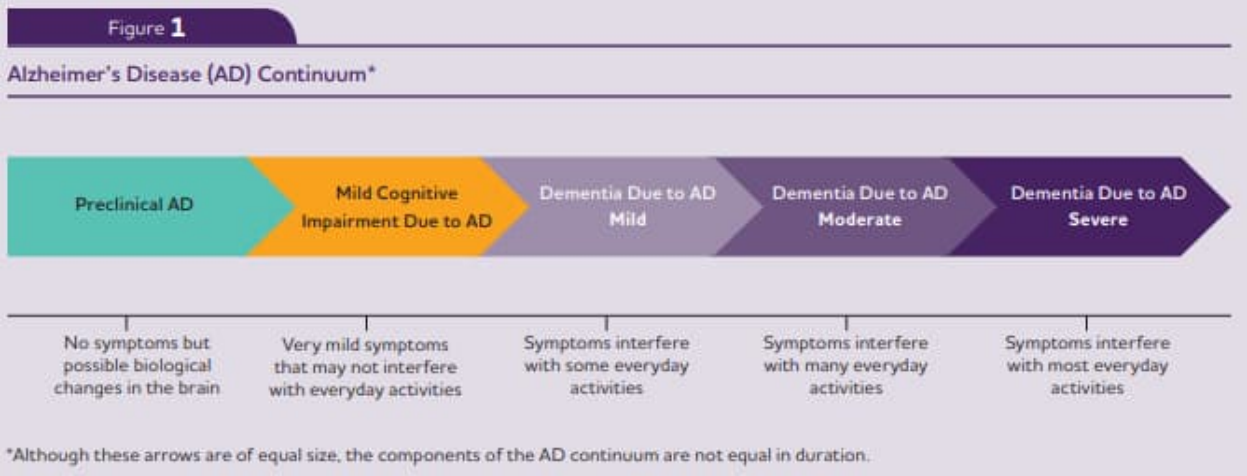
2. NEUROFIBRILLARY TANGLES (TAU TANGLES)
Accumulation of tau proteins inside neurons (called tau tangles) leading to disrupt the transport system within the neurons.

3. NEURODEGENERATION: LOSS OF SYNAPSES AND NEURONS
The neurons die because of the malfunction of the communication between neurons.

4. ATROPHY
As neurons die and connections are lost, the volume of the brain decreases.

STAGES OF ALZHEIMER'S DISEASE (AD)

Alzheimer's disease is divided into the five stages: preclinical AD, MCI (Mild Cognitive Impairment) due to AD, Mild, Moderate and Severe AD.



https://www.alz.org/media/documents/alzheimers-facts-and-figures.pdf

PRECLINICAL AD
Changes in the brain may start a decade or more before memory and other cognitive problems appear. This stage is usually not noticeable to the individual.

MILD COGNITIVE IMPAIRMENT DUE TO AD (MCI)
Some memory issues and cognitive difficulties become apparent but are not severe enough to interfere significantly with daily life. Not everyone with MCI develops Alzheimer's disease.

MILD AD
Noticeable memory loss and cognitive difficulties. Friends and family may start to notice changes.

MODERATE AD
Memory loss and confusion grow worse, and people begin to have problems recognizing family and friends.

SEVERE AD
Individuals lose the ability to respond to their environment, carry on a conversation, and eventually control movement.

Progression and symptoms can vary significantly from person to person.

LEAD (LEVERAGING EXPLORATORY ANALYSIS AND DATA IN ALZHEIMER'S DISEASE) DATA MART

The Alzheimer's disease data mart build within Roche includes demographics, clinical data, images and fluid biomarkers up to 25 years of follow-up in all stages (preclinical AD, MCI, Mild, Moderate and Severe AD) and cognitive healthy participants.
The work to collect and harmonize the data started in May 2019. The data mart follows FAIR principles.

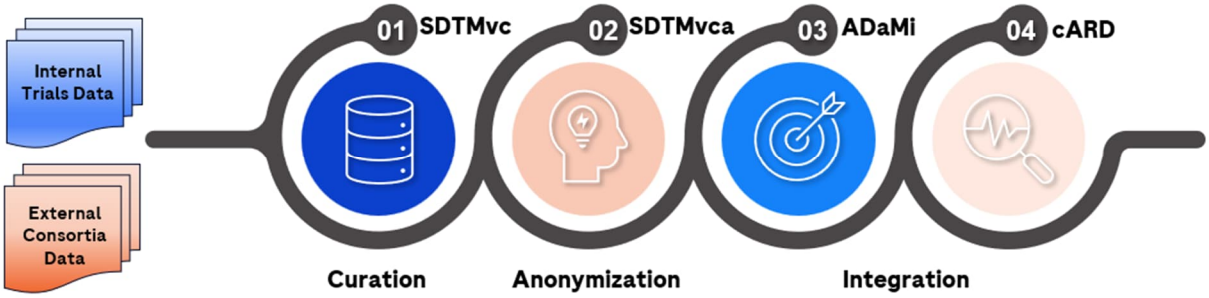
The LeAD data mart includes 35 studies, of which 14 are from in-house, with more than 75,000 patients. All Roche employees can access the data via a Data Portal or a data mart dashboard. This requires a special training upfront to handle the data, especially for the external studies. One example included in the training could be how to inform the data providers about any external publication resulting from their data and whom to cite in any presentation/publication.

FAIR

"FAIR Data Shared" is well-organized data shared for generating value and insights. The data is designed to be findable, accessible, interoperable and re-usable. FAIR principles drive mastery of information. Applying FAIR principles to how we organize, manage, and store our data makes it easier and faster for us to find, access, and transform it into actionable insights to inform drug discovery and development activities. Not just for us, as humans, but also the machines and applications that drive automated workflows and sophisticated analytics.
FAIR Data Shared is the foundation for accelerating and innovating the discovery, development and delivery of more personalized treatments for the patients. (The FAIR Guiding Principles for scientific management and stewardship by Mark D. Wilkinson, 2016)

PROCESS FLOW (DATA FLOW FROM SDTM -> ADAMI)

In Alzheimer's use case, data from external consortia and data from internal trials are received. At Roche, the SDTMvc datasets are processed and delivered based on Clinical Data Curation Guidelines and validated reference systems. These data have been curated and harmonized.
At the next step, the data is anonymized applying specific rulesets: here an identifiable person's data (personal data) is turned into data, which no longer relates to an identifiable person, thereby reducing the risk of re-identification to a negligible level (Anonymization). At the next step these SDTMvca dataset are pooled and integrated into ADaMi datasets and transformed into analysis ready datasets (cARDs). The analysis ready dataset (cARD) in LeAD is a horizontal dataset created from several ADaMi's per patient per analysis visit including the main variables for the stakeholders to do their analysis, e.g. demographic, laboratory and questionnaire data are summarized together.

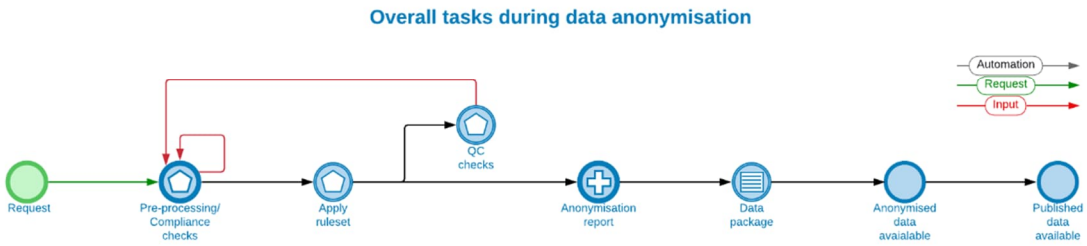


Roche has adopted a practical and relative/risk-based anonymization framework, which aims to reduce the risk of re-identification in a manner that is proportional to the data and context in which the data are shared. Use of anonymized data is best practice for secondary use of data at Roche.

The anonymization service at Roche can be automatically requested. Pre-processing compliance checks and QC checks are part of the automation process. A specified ruleset is applied and the data is anonymized.
Here a ruleset is a group of rules applied to a dataset or variable for a given data model for a given anonymization scenario. Each ruleset is based on underlying guidance papers or SOP.

An example is the variable [AGE] and the following rule: Derive a new value to keep the integrity, which means in detail: if a subject is > 89 years old then a new variable [AGE_DI] will be populated with ">89" and [AGE] will be set to blank, for ages <= 89 [AGE] will stay the same.

Process flow when using the Anonymisation service

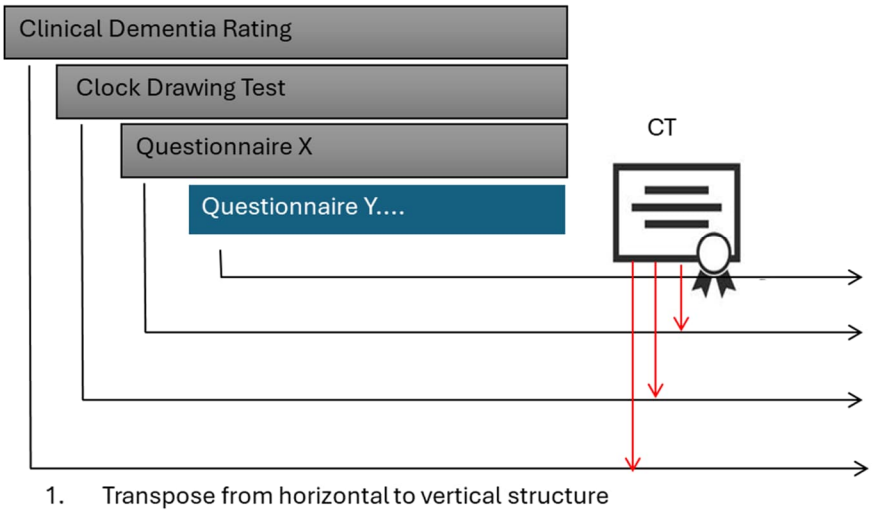


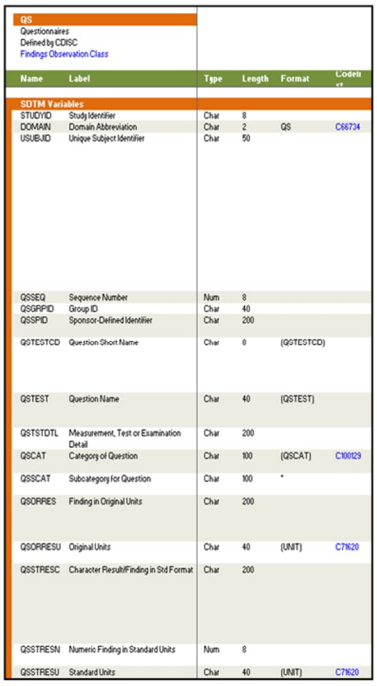
CHALLENGES

CHALLENGES IN CURATION

As described earlier, in Alzheimer's disease use case the basis are data from external studies and data from internal trials: This implies combining diverse data sources, formats and varying standards.
During curation, this data is standardized based upon CDISC SDTM implementation guidelines and according to the standard format SDTMv at Roche. Here the metadata and the context around the data is of high importance (e.g. a clear and concise documentation of the data, communication with study personal to get a better understanding of the data). As a further challenge, the interoperability across sources needs to be ensured and as per anonymization processes the data privacy and security needs to be maintained.

Example: Curation of external data - Questionnaires mapped into SDTMv domain QS





QS
Questionnaires
Defined by CDISC
Findings Observation Class

Name	Label	Type	Length	Format	Codelist
SDTM Variables					
STUDYID	Study Identifier	Char	8		
DOMAIN	Domain Abbreviation	Char	2	QS	C66734
USUBJID	Unique Subject Identifier	Char	50		
QSSEQ	Sequence Number	Num	8		
QSGRPID	Group ID	Char	40		
QSSPID	Sponsor-Defined Identifier	Char	200		
QSTESTCD	Question Short Name	Char	8	(QSTESTCD)	
QSTEST	Question Name	Char	40	(QSTEST)	
QSTSTDTL	Measurement, Test or Examination Detail	Char	200		
QSCAT	Category of Question	Char	100	(QSCAT)	C100129
QSSCAT	Subcategory for Question	Char	100	*	
QSORRES	Finding in Original Units	Char	200		
QSORRESU	Original Units	Char	40	(UNIT)	C71620
QSSTRESC	Character Result/Finding in Std Format	Char	200		
QSSTRESN	Numeric Finding in Standard Units	Num	8		
QSSTRESU	Standard Units	Char	40	(UNIT)	C71620

DIFFERENCES BETWEEN STUDY ADAM AND POOLED ADAMI

Pooled ADaMi's in the data mart should enable the consistency and the comparability of the data between the studies. Therefore, e.g. the baseline definition is the same for all studies and not according to specific study needs.
The analysis visit windowing is more general, e.g. 6-monthly and not as specified in the studies, e.g. Week 4, Week 8, …, Week 52.
Screen failure data are often not included in the study data prepared for the creation of the CSR. However, they can be important for exploratory analyses within Alzheimer's disease and are therefore included in ADaMi. For example to select biomarkers to identify participants, which can be included in future trials. For the identification of biomarkers even additional data are included after closing of a study, e.g. blood samples are re-measured.

Therefore, the results can be either different and/ or more comprehensive to the CSR read-out.

CHALLENGES IN STUDY DATA VERSUS POOLED DATA

COMPARISON OF VALUES, E.G. LABORATORY VALUES

The following example explains the difference between study data and pooled data from different studies. To unify laboratory values, e.g. the parameter Cerebrospinal fluid Amyloid Peptid Beta 40 (ng/mL), the variables laboratory method of test or examination [LBMETHOD], laboratory detailed method of test or examination [LBMTDTL], laboratory sample identifier [LBSMPLID], and reagent lot number [LBRLOT] need to be included as well to assign the parameters. This way the variables are comparable.

PARAM	PARAMCD	LBTESTCD	LBMTDTL	LBMETHOD	LBSMPLID	LBRLOT
CSF: Amyloid Peptid Beta 40 (Elecsys Batch Neurotoolkit) (ng/mL)	CFAB40	AMYLB40	ELECSYS ROCHE DIAGNOSTICS NEUROTOOLKIT	ELECTROCHEMILUMINESCENCE IMMUNOASSAY		
CSF: Amyloid Peptid Beta 40 (Elecsys at Blennow lab Batch at Screening) (ng/mL)	CFAB40E3	AMYLB40	ELECSYS ROCHE DIAGNOSTICS NEUROTOOLKIT	ELECTROCHEMILUMINESCENCE IMMUNOASSAY	ALIQUOT 1	
CSF: Amyloid Peptid Beta 40 (Elecsys at Blennow lab Batch only patients with follow-up) (ng/mL)	CFAB40E4	AMYLB40	ELECSYS ROCHE DIAGNOSTICS NEUROTOOLKIT	ELECTROCHEMILUMINESCENCE IMMUNOASSAY	ALIQUOT 2	
CSF: Amyloid Peptid Beta 40 (Lumipulse) (ng/mL)	CFAB40L	AMYLB40	LUMIPULSE FUJIREBIO	CHEMILUMINESCENT ENZYME		
CSF: Amyloid Peptid Beta 40 (Elecsys Batch Neurotoolkit) (ng/mL) (Lot K05)	CFAB40L1	AMYLB40	ELECSYS ROCHE DIAGNOSTICS NEUROTOOLKIT	ELECTROCHEMILUMINESCENCE IMMUNOASSAY		K05
CSF: Amyloid Peptid Beta 40 (Euroimmune) (ng/mL)	CFAB40M	AMYLB40	EUROIMMUNE	ELISA		
CSF: Amyloid Peptid Beta 40 (Eli Lilly Proprietary Assays) (ng/mL)	CFAB40P	AMYLB40	ELI LILLY PROPRIETARY ASSAYS	ELISA		
CSF: Amyloid Peptid Beta 40 (Elecsys Batch Neurotoolkit) (ng/mL) (Previous Lot)	CFAB40P1	AMYLB40	ELECSYS ROCHE DIAGNOSTICS NEUROTOOLKIT	ELECTROCHEMILUMINESCENCE IMMUNOASSAY		PREVIOUS
CSF: Amyloid Peptid Beta 40 (LCMS at PPD Batch) (ng/mL)	CFAB40PB	AMYLB40	INNOTEST FUJIREBIO	ELISA		

Within one study, the variable could be assigned to only one parameter, in the data mart this laboratory test includes nine different parameters, shown in detail in the table above in the parameter [PARAM] column, to make these variables comparable.

ISSUES WITH DATA: VALUES OF VARIABLES

In study data values of single variables are unique and clear, within pooled data it could lead to inconsistencies and/or misunderstandings, e.g. for the result of the PET (Position Emission Tomography) Visual Read [PVISREAD]. The PET Visual Read refers to the interpretation of PET scans to identify patterns showing signs of Alzheimer's disease. Positive means, that the patient does have amyloid plaques in the brain, and negative that not. The answers of "YES" / "NO" needs to be harmonized into "POSITIVE" / "NEGATIVE" to be consistent with the studies already included.

PVISREAD	Subtotal
	NO
	YES
	Ambiguous/indeterminate
	Negative
	Positive

Table 1: PET Visual Read [PVISREAD]

VARIABLE IDENTIFIERS

Generating mean line graphs of numeric values can show inconsistencies within pooled studies. The following example shows a mixing up of a test code, which is the reason of the straight line at the bottom of the graph:
The variable of ADCS-ADL: "Dispose of Garbage Usual Performance" [DLGARB1] has the three different codes 1 (with physical help), 2 (with supervision), and 3 (without supervision or help) and is coded to "ADL0114A" for Questionnaire Test or Examination Short Name [QSTESTCD]. In one study it was unfortunately coded to "ADL0114A" for the question, if the subject did dispose garbage or litter, which is answered with 0 (no) or 1 (yes).

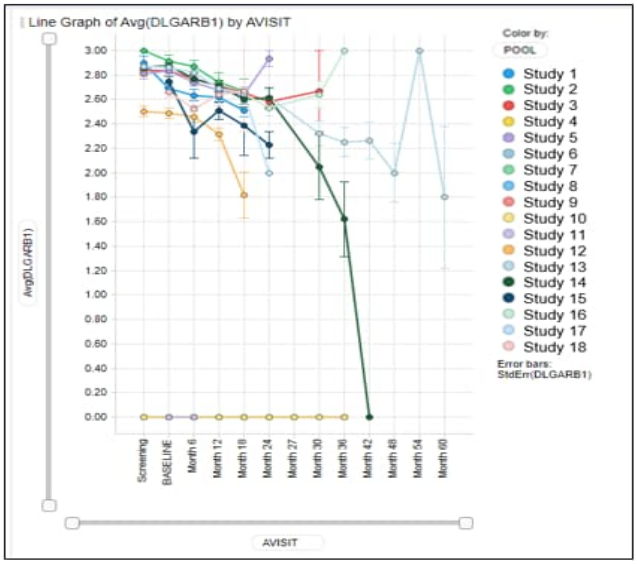


Figure 1: Plot of Activity of Daily Living – Garbage Performance by Analysis Visit

ISSUES WITH DATA: VARIABLE UNITS

Another example of strange results lead to a new data transfer, since the values together with the units were wrong in the source file. The values shown in the plot of the average for the two studies included in addition were that high, that the mean values from all other studies are shown in the plot as one straight line around zero.

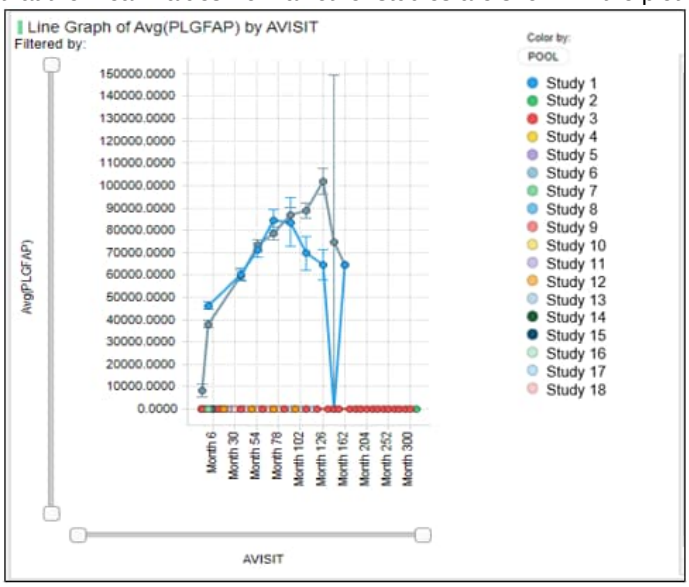


Figure 2: Plot of mean plasma glial fibrillary acidic protein (GFAP) (pg/mL) by analysis visit

ISSUES WITH DATA: VOLUMETRIC MRI DATA

Comparing values between studies to check if they make sense. The reason that values do not match can also have a not so easy to spot reason in the source data. In the following example the values were not calculated correctly, instead of the sum, the average was calculated of the left and the right hippocampal volume of the brain.

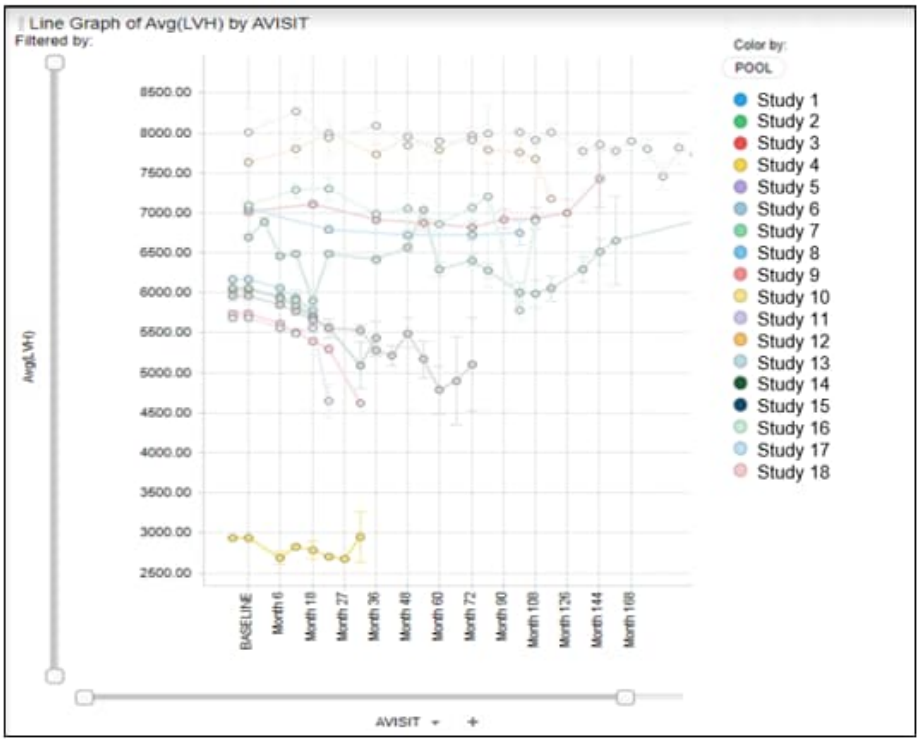


Figure 3: Plot of mean hippocampal volume (MRI) by analysis visit

UPDATES OF STANDARDS WITHIN COMPANY

Over Time, standards develop further. For the laboratory vendor name [LBNAM] it changed from short to long names, e.g. “COVA” to “Covance”, and from upper case to lower case, e.g. “COVANCE” to “Covance”. This could lead to updates in parameter settings or in programming.

UPDATES IN SDTM STANDARDS

The mapping of the variable ApoE Status was used in several SDTMs in the past. The ApoE gene is important for the production of a protein that is crucial in the metabolism of fats in the body. The ε4 allele is associated with an increased risk of developing late-onset Alzheimer’s disease. An individual can have zero, one, or two ε4 alleles. This variable is captured so far in four different SDTM domains: SC, ZB, PF and GF in the curated studies:

SDTM.SC (4 studies)
Stratification variable [SCORRES] = 0e4, 1e4, 2e4

SDTM.ZB (4 studies)
[ZB.ZBORRES] where [ZB.ZBTESTCD] = “ALLELE”

SDTM.PF
[PF.PFSTRAIN] where [PF.PFTESTCD] = “ALLELE” and [PF.PFGENRI] = “APOE”

SDTM.GF (newest)
[GF.GFORRES] where [GF.GFTESTCD] = “ALLELE” and [GF.GFSYM] = “APOE”

The studies are curated separately and according to the standards at the time included in the data mart. The data are then pooled together and integrated in the ADaMi’s, in this example within the parameter APOE4, which contains 0e4, 1e4 or 2e4.

FEEDBACK OF USERS

More than 50 users did use the generated Alzheimer’s data mart across the organization, e.g. biostatistics, genetics

group, diagnostics, scientists, patient reported outcomes, biomarker development.

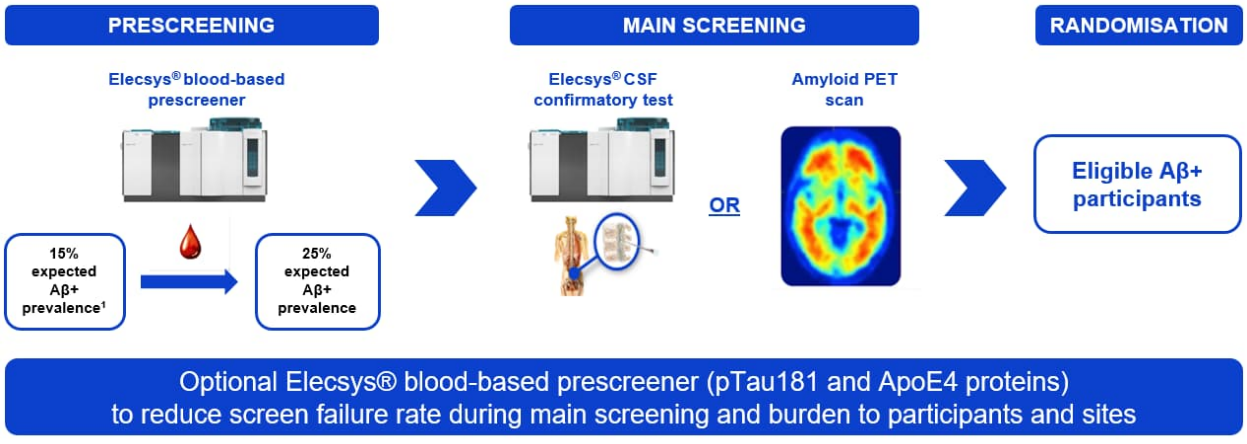
There is positive feedback from users, e.g. that the “LeAD Alzheimer data mart was helpful and the instructions how to access and download the data simple, accurate and concise. This made a complicated process a lot easier, and expedited the entire activity.” (Stephen). “It was very helpful, to pool the in-house studies ourselves would have been much more time consuming” (Sandro). “It was easy to answer a question, even in between meetings” (Brad). The data mart was a trigger for the “Exchange of statistical methods between teams” (Kirill).

USAGE

The pooled LeAD Alzheimer data mart data were already used for a variety of use cases, e.g. for clinical trial design, digital platform for diagnosis, access and pricing strategy, prognostic models and biomarker analyses.
For clinical trial design, the data were explored for the definition of inclusion and exclusion criteria, which questionnaire data should be included in the trial. Data were used to for Alzheimer’s disease progression models.

BLOOD-BASED BIOMARKERS IN ALZHEIMER’S DISEASE

One other example was the use of blood-based biomarker prescreening in Alzheimer’s disease for a secondary prevention study called Skyline. Different plasma samples from different studies of the pooled Alzheimer data mart were re-analyzed to find a plasma biomarker to rule out patients, to reduce the screen failure rate during the main screening. For identifying eligible patients in a clinical trial for Alzheimer’s disease either a CSF test or an amyloid PET scan is needed, which is a high burden for patients, cost-intensive plus time consuming. The blood-based prescreener (pTau181 and ApoE4 proteins) were identified. Afterwards, the correctness of the prescreening was confirmed with the study data.



Elecsys® blood-based prescreener is investigational and availability depends on local country / regulatory approval.
Elecsys® β-Amyloid (1-42) CSF and Elecsys® Phospho-Tau (181P) CSF are currently available in CE-mark accepting countries.
Aβ, beta-amyloid; ApoE4, apolipoprotein E ε4; CSF, cerebrospinal fluid; pTau, phosphorylated tau.
1. Sperling RA, et al. JAMA Neurol 2020;77:735–745.

Bittner et al. ADPD conference presentation 2024

CONCLUSION

FAIR data is making secondary use of the data much easier. In addition, keep the needs of the stakeholders in mind when preparing the data. Working with the data needs an understanding of the disease and the data.
Always think of secondary use of study data in the process of generating data. Furthermore, good documentation practice is the key.

REFERENCES

Lancet Neurology vol. 12; Jack CR Jr, et.al, pp207-216;2013
Alzheimer's Disease Facts and Figures: https://www.alz.org/media/documents/alzheimers-facts-and-figures.pdf
Sperling RA, et al. JAMA Neurol 2020 ; 77 :735-745
Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).
Galasko D, Bennett D, Sano M, et al. An inventory to assess activities of daily living for clinical trials in Alzheimer's disease. The Alzheimer's Disease Cooperative Study. Alzheimer Dis Assoc Disord. 1997;11 Suppl 2:S33-9 (PubMed abstract)
Tables and Figures created with TIBCO Spotfire®
ChatGPT
MS Powerpoint

RECOMMENDED READING

Alzheimer's disease Facts and Figures: https://www.alz.org/media/documents/alzheimers-facts-and-figures.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Michaela Mertes
Company: F. Hoffmann-La Roche Ltd
Email:michaela.mertes@roche.com

Author Name: Iska Gans
Company: F. Hoffmann-La Roche Ltd
Email:iska.gans@roche.com