Paper RE03

Using the USDM and Ensuring Governance and Traceability to unlock the potential of Real-World Data sources for regulatory submissions

Berber Snoeijer, ClinLine, Leiderdorp, The Netherlands
Alistair Dootson, EQTY Life Sciences, Manchester, United Kingdom

ABSTRACT

The goal of real-world data (RWD) is to be used as evidence (RWE) in regulatory submissions. The governance and traceability of this RWD are paramount, ensuring integrity, reliability, and compliance for regulatory submission. The use case described in this paper shows how the eligibility criteria of the Lilly LZZT study are populated in the CDISC unified study definitions model (USDM) using syntax template capabilities and how this drives the selection of real-world data for an external control arm. Furthermore, this paper explores challenges and best practices associated with governing RWD throughout its lifecycle, from collection at the source to submission, providing regulators proof of source reliability without regulators needing to access third-party source data.

Building upon the ideas of multiple sponsors and regulators through working group collaborations, we developed a method to demonstrate principles and best practices in data governance, quality, and traceability. Utilizing data standards, code, and a software-agnostic tool that created cryptographic certificates for each data set and then weaved them together to form a verifiable data mesh, an Alzheimer cohort was created for a mock submission. Based on patients selected from FHIR electronic health record (EHR) resources, data was traced through its transformation to OMOP+ then anonymized, simulating a data vendor process. Subsequently, fit-for-purpose selection and then transformation to SDTM was conducted simulating sponsor/CRO processes.

Through this paper, we will demonstrate how the USDM can inform this process, how data reliability is now verifiable by reviewers and how new open-source cryptographic technologies automate data lineage and contemporaneously document adherence to governance controls, providing the highest confidence for data and process reliability.

INTRODUCTION

Regulatory guidelines indicate that reliability of real-world data should be demonstrated and as such data lineage is more and more a term that is used for this concept. Data lineage refers to the process of tracking how data is **generated**, **transformed**, **transmitted** and **used** across a system over time. It documents data's origins, transformations and movements, providing detailed visibility into its life cycle. This process simplifies the identification of errors in data analytics workflows, by enabling users to **trace issues back to their root causes** (Wikipedia).

The EMA Data Quality Framework indicates that the reliability of clinical data fundamentally depends on the systems and processes in place for the primary collection of data and its processing (1). In alignment with that, the FDA RWD guideline on assessing electronic Health Records and medical claims data to support regulatory decision making, indicates that sponsors should address the procedures used to ensure completeness and accuracy of study data, as well as processes for data accrual, curation and transformation over the data life cycle (2). Ensuring this data lineage from real-world data source to submission is however challenging given the diversity of data sources, documentations, and data transformations.

By focusing on a specific RWE use-case, we demonstrated how the lineage process can be informed by the USDM, how study design eligibility criteria can be directly used to automate source data selection and how cryptographic certificates mathematically prove the integrity of data for regulatory submission. With verification of data, code or actions down to the hardware level, this method proves a sponsor's compliance to regulations and supports transparency and explainability of complex data workflows.

USE CASE DESCRIPTION

For this use-case we used mimicked patient data from an EHR being licensed to create submission-ready datasets. We used FHIR patient files generated from Synthea data (3) and mapped them to structures similar to OMOP (4) for further selection and fit-for-purpose assessments.
In addition, we defined a hypothetical external control arm in USDM format (5) to facilitate the automated selection and fit-for purpose assessment of this standardized real-world data. Finally, the data of the selected patients was transformed to SDTM for study analysis and submission purposes.

USING OMOP AS A CDM FOR PATIENT AND DATA SELECTION

As described in the FDA guideline for Assessing Electronic Health Records (EHR) and Claims data (2), it is important to standardize data from a variety of EHR data sources into a common format to ensure interoperability across all sites providing data. Although FHIR is a good standard for data exchange, the FHIR format is less suitable for patient and data selection. Therefore, for this purpose we used the OMOP (4) observational CDM to transform the data for these purposes.

This approach was favored above the direct transformation from FHIR to SDTM for multiple reasons.

1) Transformation directly from FHIR to SDTM implies cumbersome transformations of data needed for selection and data source fitness for purpose evaluation but not for submission. It would also include the transformation of patient data that are not included in the study to be able to assess whether they are fit or not.
2) Although beneficial for the selection of patients, electronic health data includes several variables that are not included in SDTM. While, for traceability purposes, OMOP includes source as well as standardized values which improves the ability of selecting and doing bias assessments regarding source collection practices or data standardization practices. This also aids the regulatory requirement to have adequate processes in place to ensure confidence in the resultant data like electronic documentation.
3) OMOP structure and comparable data vendor structures are fairly basic and once understood, easy to use for selecting data and patients.
4) Aligning to observational research requirements, data vendors often utilize the OMOP common data model (CDM) or similar structures. Therefore, mapping to OMOP exemplifies the transformation of source Electronic Health Records (EHRs) to a CDM, enabling the execution of identical queries (without any substantial modifications) on multiple datasets.

USE CASE PROCESS

For this use case we took a step by step approach.

Step 1: Automated patient selection from the source using USDM
Indication disease codes that are used in the source are specified in USDM. Based on these codes the patient FHIR data files were automatically selected from the source data.

Step 2: Read and standardized FHIR JSON data
At the location of the data provider, the FHIR data of selected patients will be extracted and mapped to an OMOP-like structure for cohort building, eligibility evaluation and fit for purpose assessments.

Step 3: Anonymization
The data in these files will be further deidentified using hashing techniques and date shifting. When needed, this technique can be utilized again on the source data to subsequently get the same anonymized ids and thus enable traceability back to the source.

Step 4: Automated selection of Eligible Patients
Based on eligibility specifications in USDM format the eligibility of patients is evaluated and a corresponding attrition diagram is prepared.

Step 5: Fit for Purpose assessment
Further fit for purpose assessments is based on evaluation of availability of data for eligible patients and sensitivity assessments. In addition, the sponsor can then apply additional filtering based on data inspections, apply imputations, and perform matching with the interventional arm.

Step 6: SDTM mapping
The selected patients and corresponding data is mapped to SDTM datasets using standardized mapping processes.

Step 7: Proof auditable Traceability
The output and mapping information is used to reproduce the source data for verification purposes.

STANDARDS AND PROCESS

USDM SPECIFICATIONS

For this use case we used USDM v3.0 (5) API specifications aligning with the LZZT trial. The study design stored in this USDM API format, drove a number of steps in the use case:

- The study indication, stored in the USDM indication class was used to drive the automated selection of patients (Step 1) for the cohort. The USDM allows for adding one or more indication codes aligning to the dictionary used in the EHR source. In our case, the source used SNOMED codes. Therefore the following information was stored in the USDM indication class addressing the Alzheimer indication:

Name	Label	Codes.code	Codes. codeSystem	Codes. CodeSystemVersion	Codes.decode
IND2	Alzheimer's disease	26929004	SNOMED	2024-02-01	Alzheimer's disease
		23026500	SNOMED	2024-02-01	Familial Alzheimer's disease

- The eligibility criteria are stored in the USDM eligibility Criterion class. The USDM syntax template tagging mechanism applicable to this class, enabled us to refer to data elements stored elsewhere in the USDM or outside the USDM. The colored items in the LZZT example below shows what items we tagged for this use case. This information was used to drive the automated selection of eligible patients (Step 4)

category	identifier	Eligibility Criteria with resolved tags (categorized)
Inclusion Criteria	01	Males and postmenopausal females at least 50.0 Year of age.
Inclusion Criteria	02	Alzheimer's disease as defined by National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disorders Association (ADRDA) guidelines (Attachment LZZT.7).
Inclusion Criteria	03	MMSE score of 10 to 23.
Inclusion Criteria	04	Hachinski Ischemic Scale score of ≤4 (Attachment LZZT.8).
Inclusion Criteria	05	CNS imaging (CT scan or MRI of brain) compatible with AD within within the past 1 year. The following findings are incompatible with AD: a. Large vessel stroke 1. Any definite area of encephalomalacia consistent with ischemic necrosis in any cerebral artery territory. 2. Large, confluent areas of encephalomalacia in parieto-occipital or frontal regions consistent with watershed infarcts. The above are exclusionary. Exceptions are made for small areas of cortical asymmetry which may represent a small cortical stroke or a focal area of atrophy provided there is no abnormal signal intensity in the immediately underlying parenchyma. Only one such questionable area allowed per scan, and size is restricted to ≤1cm in frontal/parietal/temporal cortices and ≤2 cm in occipital cortex.

The colors indicate to what part of the USDM the tagged items are referencing. Dark Green: population characteristics, Red: Assessments stored as biomedical concepts, Light Green: timeline items. Gray: tags that are used for automation but not for referencing to other parts of the data model.

- The timeline and corresponding timing information stored in the USDM can be used to select the source visit data that correspond to the study schedule. The timing information, including the corresponding allowed window is stored in ISO 8601 format which in our use case could drive the selection of usable encounters for the external control arm. The fit for purpose assessments (Step 5) were based on availability of visits and assessments. Due to time limitations, we did not include that automation yet in this use case.
- Planned assessments like lab tests are linked to the corresponding planned activities in the USDM timeline. These assessments were stored as biomedical concepts. For this use case, we added alias codes to these biomedical concepts that matched the codes in the source. This allowed us to automatically select and evaluate the eligibility (Step 4) and the availability (Step 5) of assessments in the EHR source data. Furthermore, as both the SDTM code as well as the aliascode matching the source were stored in the biomedical concept, that allows for automated SDTM mapping (Step 6). In our use case, we showed the automated evaluation of eligibility criteria. The automated mapping steps will be show in a future use case.

The table below shows how the Biomedical concepts of our use case were stored in the USDM. It shows that for each biomedical concept multiple alias codes were specified which aligned to the varied specifications in real-world data sources (SNOMED and LOINC).

MYBCS — Table rows: 20 | Columns: 19 of 19

Enter expression

	tag	BCId	AliasId	code	codeSyst...	codeSystemVersion	decode	instanceType
1	BC1	BiomedicalConcept_x001	Code_x003	273617000	SNOMED	2023-07-31	Mini-mental state examination	Code
2	BC1	BiomedicalConcept_x001	Code_x004	72107-6	LOINC	2.76	Mini-mental state examination [MMSE]	Code
3	BC1	BiomedicalConcept_x001	Code_x091	273619002	SNOMED	2023-07-31	Mini-mental state examination	Code
4	BC1	BiomedicalConcept_x001	Code_x092	72106-8	LOINC	2.76	Total score [MMSE]	Code
5	BC3	BiomedicalConcept_29	Code_x008	45896001	SNOMED	2023-07-31	Aspartate aminotransferase measurement	Code
6	BC3	BiomedicalConcept_29	Code_x009	1920-8	LOINC	2.76	Aspartate aminotransferase [Enzymatic ...	Code
7	BC4	BiomedicalConcept_8	Code_x011	34608000	SNOMED	2023-07-31	Alanine aminotransferase measurement	Code
8	BC4	BiomedicalConcept_8	Code_x012	1742-6	LOINC	2.76	Unit of Catalytic Activity Concentration	Code
9	BC5	BiomedicalConcept_30	Code_x013	70901006	SNOMED	2023-07-31	Creatinine measurement	Code
10	BC5	BiomedicalConcept_30	Code_x014	14682-9	LOINC	2.76	Creatinine [Moles/volume] in Serum or ...	Code
11	BC5	BiomedicalConcept_30	Code_x097	38483-4	LOINC	2.76	Creatinine [Moles/volume] in Blood	Code
12	BC6	BiomedicalConcept_x002	Code_x016	441689006	SNOMED	2023-07-31	Measurement of total hemoglobin conc...	Code
13	BC6	BiomedicalConcept_x002	Code_x017	718-7	LOINC	2.76	Hemoglobin [Mass/volume] in Blood	Code
14	BC7	BiomedicalConcept_31	Code_x013	59573005	SNOMED	2023-07-31	Potassium measurement	Code
15	BC7	BiomedicalConcept_31	Code_x014	2823-3	LOINC	2.76	Potassium [Moles/volume] in Serum or ...	Code
16	BC8	BiomedicalConcept_32	Code_x013	25197003	SNOMED	2023-07-31	Sodium measurement	Code
17	BC8	BiomedicalConcept_32	Code_x014	2951-2	LOINC	2.76	Sodium [Moles/volume] in Serum or Pla...	Code
18	BC9	BiomedicalConcept_x003	Code_x013	71878006	SNOMED	2023-07-31	Calcium measurement	Code
19	BC9	BiomedicalConcept_x003	Code_x014	2000-8	LOINC	2.76	Calcium [Moles/volume] in Serum or Pl...	Code
20	BC9	BiomedicalConcept_x003	Code_x095	38483-4	LOINC	2.76	Calcium [Moles/volume] in Blood	Code

- Finally, the general study design information like study types, intervention information, objectives and endpoints and more can be used to automatically create SDTM trial design domains. The corresponding mapping information is available in the USDM IG. This will be shown in a future use case.

DATA FLOW-TRANSFORMATION FHIR TO OMOP

For the transformation of FHIR to the OMOP-like tabular structure, we tried a number of open-source and investigated proprietary AI based abstraction tools. However, either a lot of mapping information still had to be provided to those open source tools available, not all information was automatically mapped, or the mapping tools themselves were a black-box, causing issues with the backward compatibility. Therefore, we chose to use the SAS® JSON option in the LIBNAME statement. This ensured that all the information from the source FHIR files was extracted and we could select what we needed.

In addition to the JSON files, a number of mapping files were used to allow for automated transformation from FHIR to OMOP and from source dictionaries to standardized dictionaries. The full path from source towards CDISC was documented to show the full trail from source to submission values and to avoid any re-mapping issue (See Mapping example below).

	Synthea / FHIR		OMOP				CDISC		
Codelist	Value	Code	concept_id	Code	Name	Vocabulary	Codelist	Codelist code	CT
Employment	full-time	0	4053118	160903007	Full-time employment	SNOMED	C111108	C52658	FULL-TIME
Employment	part-time	1	4059634	160904001	Part-time employment	SNOMED	C111108	C75562	PART-TIME
Employment	unemployed	2	4251171	73438004	Unemployed	SNOMED	C111108	C75563	NOT EMPLOYED
Employment	retired	3	4022069	105493001	Retired	SNOMED			

Not all data from the original FHIR files was mapped. Because of proportionality reasons, data that is not necessary for further processing or traceability should not be included. Therefore, before mapping, the data vendor needs to assess what information might be relevant for clinical research, representative of both negative and positive lineage. As a result, for this proof of concept, the following FHIR resource types were included in the mapping:

- Patient
- Observation
- Condition
- AllergyIntolerance
- MedicationRequest
- Immunization
- Procedure
- Encounter
- Location

**Note in actual practice a data vendor might want to address more resources to ensure consistency and drug accountability checks.*

The created OMOP like datasets including supplemental variables to hold the received FHIR content as is, were based on the FHIR information available and the potential needs for clinical research. In this proof of concept, the following OMOP like datasets were created:

- Person
- Death
- Visit_occurrence
- Condition_occurrence
- Procedure_occurrence
- Measurement
- Observation
- Fact_relationship

TRANSFORMATIONS FROM OMOP TO SDTM

Like for the transformation from FHIR to OMOP like datasets, for each resulting SDTM dataset a mapping file was created indicating the exact transformations needed to convert to SDTM data sets. These transformation files, along with the dictionary mapping file, were used to automate the mapping from OMOP to SDTM. In our proof of concept, we included the mapping to the following SDTM domains that served as a good representation of the full set of SDTM domains required for submission for a safety control arm:

- DM
- SV
- VS
- SC
- AE
- MH
- CM
- LB
- QS
- Multiple supplemental datasets

Important to note is that there were two special cases where mapping of real-world data to SDTM submission datasets is generally more challenging and requires pre-configured governance. This was the case in determining whether conditions should be mapped to the Adverse event (AE) or Medical History (MH) domain. Our decision was based on the index date. Hence, all conditions before the index date were mapped to MH, while all conditions after the index date were mapped to AE. In case relatedness information is available in the source, we could add that to the data. When a MedDra licence is available, OMOP includes a mapping dictionary from SNOMED to MedDra which allows for the mapping of most conditions directly to MedDra. For the Concomitant Medication (CM) mapping such a mapping is not available. However, this can be partly handled by directly matching substance and drug names to the corresponding WHO drug codes needed for submission. In addition, the OMOP concept_relationship and drug_strength databases can aid the mapping of other CM variables considerably.

DATA ANONYMIZATION

To ensure adherence to patient privacy regulations, the data was anonymized with proof at the processor level of transformations and anonymization methods, which documented the R factor below .09 for legal standing. Before being then utilized in downstream clinical study purposes, the licensing agreement for each data set was verified. For regulatory purposes, traceability back to the initial source record needs to be maintained. To address this lineage to provenance issue along with the complexity of regulators accessing third-party data, the following was accounted for in this proof of concept:

- Cryptographic certifications were used as deidentification measures for patient, location and provider IDs. This enabled tracing from source to end but not the other way around. Hence, one needs to know the patient IDs in the source to restore the connection to the deidentified data.
- Date shifting was applied by a maximum of 2 years. Shifts were applied consistently per patient to keep the consistency of the data within a patient.
- Sensitive deidentifying observations like questions on migration status, jail, address and income question were deleted.

Not conducted in the reduced scope, in practice it is recommended to review the full list of assessments included in a source, gauging sensitivity and documenting its inclusion or exclusion for clinical study purposes, which then is traceable. Furthermore, per variable, the distribution of values within and between subjects should be assessed to evaluate whether this poses any deidentification risk and ensure data is not duplicated.

FEASIBILITY AND COHORT SELECTION

As indicated above, the indication (Alzheimer's disease) and eligibility stored in a standardized format in the USDM drove the automated selection of patients from the source. The example below shows how the selection based on the criteria indicated in time when a patient was eligible. Patient 2159 in this example

- turned 50 on 2019-02-02 (the first criterium is met, however, the next criterium for MMSE is still > 23)
- MMSE is 15.6 on 2019-06-24 (now the second criterium, for MMSE, is also met)
- ALAT was also within the limits on 2019-03-25 and other assessment criteria as well.

Therefore the patient was evaluated to be eligible and the corresponding index date was set to 2019-06-24. This is the date that all criteria were met.

Of course using older lab results without timing limitations for the assessment might be questionable. Therefore, it is advisable to add timing limits to the real-world eligibility criteria. Furthermore, it needs to be addressed what to do when certain variables for eligibility are not available for a patient. Excluding these patients might not be the appropriate way as discussed in the FDA guideline (2). For our use case, we assumed that the missing values were within the limits.

ALLSELLIM — Table rows: 18614 | Columns: 21 of 26 | Rows 18401 to 18600

Enter expression

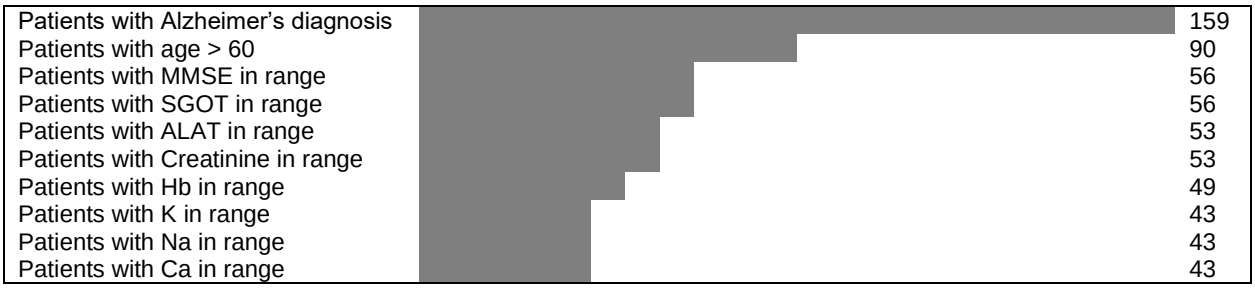
	new_patno	date	source_value	value_...	unit_sourc...	Age1	BCEvalUnit	BC1	BC2	BC3	BC4	BC5
18552	49 F	2159	2018-03-25	Creatinine [Mass/volum...	3.1037	mg/dL	0	Micromole per Liter	.	1	.	0	.
18553	49 F	2159	2018-03-25	Calcium [Mass/volume] i...	9.37	mg/dL	0	Millimole per Liter	.	1	.	0	.
18554	49 F	2159	2018-08-04	Familial Alzheimer's dis...	.		0		.	1	.	0	.
18555	49 F	2159	2018-08-04	Total score [MMSE]	24.716		0		0	1	.	0	.
18556	49 F	2159	2018-08-04	Creatinine [Mass/volum...	1.14	mg/dL	0	Micromole per Liter	0	1	.	0	.
18557	49 F	2159	2018-08-04	Calcium [Mass/volume] i...	8.88	mg/dL	0	Millimole per Liter	0	1	.	0	.
18558	50 F	2159	2019-02-02	Creatinine [Mass/volum...	0.82	mg/dL	1	Micromole per Liter	0	1	.	0	.
18559	50 F	2159	2019-02-02	Calcium [Mass/volume] i...	9.79	mg/dL	1	Millimole per Liter	0	1	.	0	.
18560	50 F	2159	2019-03-25	Aspartate aminotransfer...	35.592	U/L	1	Microkatal per Liter	0	1	.	0	.
18561	50 F	2159	2019-03-25	Alanine aminotransferas...	20.573	U/L	1	Microkatal per Liter	0	1	.	1	.
18562	50 F	2159	2019-03-25	Creatinine [Mass/volum...	2.8363	mg/dL	1	Micromole per Liter	0	1	.	1	.
18563	50 F	2159	2019-03-25	Calcium [Mass/volume] i...	8.9	mg/dL	1	Millimole per Liter	0	1	.	1	.
18564	50 F	2159	2019-06-24	Total score [MMSE]	15.634		1		1	1	.	1	.
18565	51 F	2159	2020-02-08	Creatinine [Mass/volum...	0.81	mg/dL	1	Micromole per Liter	1	1	.	1	.
18566	51 F	2159	2020-02-08	Calcium [Mass/volume] i...	9.71	mg/dL	1	Millimole per Liter	1	1	.	1	.
18567	51 F	2159	2020-03-24	Aspartate aminotransfer...	9.5298	U/L	1	Microkatal per Liter	1	1	.	1	.
18568	51 F	2159	2020-03-24	Alanine aminotransferas...	28.77	U/L	1	Microkatal per Liter	1	1	.	1	.
18569	51 F	2159	2020-03-24	Creatinine [Mass/volum...	3.1922	mg/dL	1	Micromole per Liter	1	1	.	1	.
18570	51 F	2159	2020-03-24	Calcium [Mass/volume] i...	9.02	mg/dL	1	Millimole per Liter	1	1	.	1	.
18571	51 F	2159	2020-05-01	Aspartate aminotransfer...	26.252	U/L	1	Microkatal per Liter	1	1	.	1	.
18572	51 F	2159	2020-05-01	Alanine aminotransferas...	49.013	U/L	1	Microkatal per Liter	1	1	.	1	.

For our proof of concept we selected 159 patients from the EHR source with the Alzheimer's disease indications. Out of these patients, 85 met all the criteria that we included in the evaluation (See the attrition diagram below)

Attrition diagram based on original criteria

To simulate the effect of changes in the criteria, we changed the population age criterium in the USDM, to show how this change influenced the attrition. This resulted in a significant reduction of eligible patients as can be seen from the diagram below. Since all selections are automated, the effect of a change in criteria could be shown in seconds which allows for a much more flexible design process.

Attrition diagram based on adjusted age criterium



INDEX DATES AND WINDOWING

A specific and difficult challenge when designing RWE external control arms for clinical trials is specifying the index date (also called *time zero* or *zero time*), which is the start of the observation period for assessing endpoints [6]. In this proof of concept, we defined the index data as the first visit date that all criteria were met. All calculated visit days and the corresponding windowing were defined based on this index date.

To align with the interventional trial arm the following windowing was applied to the real-world source data based on the index date.

Interventional trial visit day			Day calculated based on index date	
Visit	**Week**	**Planned visit Day**	**WindowLower Day**	**WindowUpper Day**
1	-2	-14	-180	-1
2	-0.3	-2	**DO NOT**	**INCLUDE**
3	0	0	**INDEX DATE**	7
4	2	14	7	20
5	4	28	21	34
7	6	42	35	48
8	8	56	49	69
9	12	84	70	97
10	16	112	98	125
11	20	140	126	153
12	24	168	154	174
13	26	182	175	188

Gaps between the windows were avoided to make sure all data within 6 months after the index date was included. Since this is real-world data, a patient could have more EHR visits within a window. All data of the multiple EHR visits are included in the final datasets to enable the use of the correct value for a specific assessment. For each measurement, we used the non-missing observation closest to the planned visit date for analysis.

GOVERNANCE AND DATA LINEAGE

GOVERNANCE

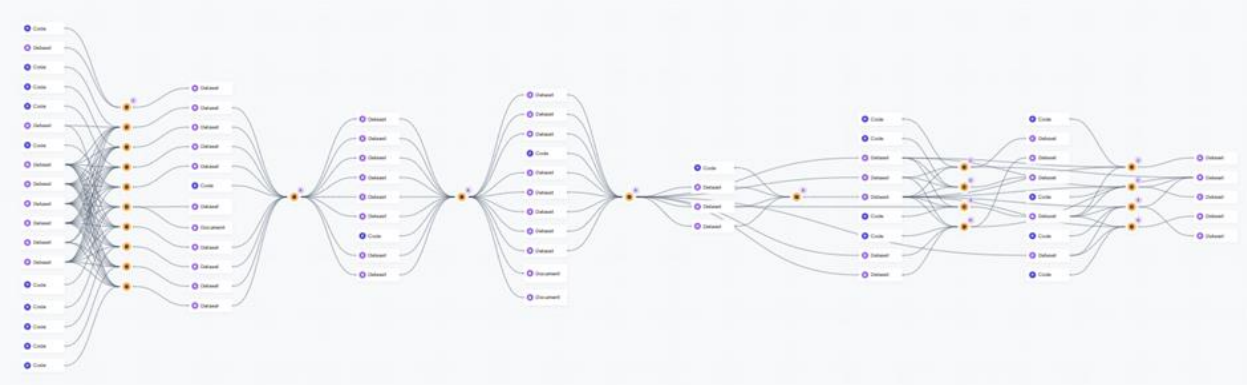
For each step in the process, the governance policies were defined to describe the requirements for provenance, traceability and quality control. Cryptographic integrity was applied to verify whether processes were followed from data source to SDTM. This ensures authenticated proof of data and processes across organizations and systems, enabling complete lineage, unified governance, and granular traceability.
The following governance policies were included in the lineage process:

Policy	Policy Rules	Policy Criterion
RWD licensing: Select data at source and transfer	(1) Verify transfer contract in place and signed (2) Verify correct run of code to select Alzheimer patients and approval by data owner (3) Verify and encrypted transfer via secured channels to data vendor (4) Verify data deletion at data vendor site within licensing period	(1) Contract includes definition of cohort including selection specifications (ie indication codes) (1) Legal review of contract (1) Signature by EHR data owner / Data vendor (2) R program to select patients based on batch of Json files in version controlled Github (2) QC of R program run by second programmer (2) Review and approval of output by EHR data owner (3) Apply encryption before sending (3) Verify secure channels/method used for sending data (acc to contract) (3) Verify data unchanged after sending to data vendor (4) Verify original (non-anonymized) data is deleted at data vendor site within 1 year of contract date
Transformation & mapping policy 1 (FHIR-OMOP)	(1) Mapping input created as input (2) Main programmer creates mapping programs from FHIR to OMOP+ (3) QC procedure	(1) Mapping files created directly linking from FHIR content to OMOP+ (1) Mapping files transferred to version controlled GitHub (2) Follow good programming guidelines (2) Input of latest version of mapping files and OMOP concept library (2) Validate against mapping specifications (2) Verify correct run of Fhir to OMOP+ programs (2) Notify when programs are ready (3) Verify alignment with OMOP concept library (3) Verify all patients from source are included (3) Verify correct run of FHIR to OMOP+ programs (3) Verify no data is changed by mapping backwards (3) Verify that programs are both signed off by programmer and QC
Anonymize Patient Data	(1) Define scope of anonymization (2) Perform anonymization according to scope (3) Verify anonymization according to scope (4) Verify secured storage of deidentification information	(1) Define shifting of dates and time variables (1) Define anoymization of location variables (1) Describe rationale for not anomyzing of certain variable for regulatory purposes (sex, age, race, country ethnic etc) (1) Define removal of potential sensitive information (pregnancy, mental health, family conditions etc) (2) Apply encryption of ID variables (2) Remove variables identified as sensitive (2) Apply shifting of date variables (3) Verify data unchanged after sending to data vendor (3) Prove Compliance with Privacy Regulations (SHA-256) (4) Compliance with Cryptographic Standards (4) Data Integrity and Consistency Checks (4) Audit and verification (SHA-256)

Data cohort selection and transfer	(1) Verify ethical approval for sponsor request is in place (2) Verify selection based on Sponsor requirements and approval (attrition diagram) (3) Verify encrypted / unchanged transfer of data to Sponsor	(1) Contract includes (ref to) data cohort requirements, ethical review requirements and proportionality requirements (1) Signatures of sponsor and data vendor on contract (1) Ethical approval in place (2) Selection specifications based on codes provided by sponsor and aligning with contract definitions (2) Program to select the patients based on criteria (2) QC of program to select patients based on criteria (3) Apply encryption before sending (3) Verify secure channels/method used for sending data (acc to contract) (3) Verify data unchanged after sending to data vendor
Fit for purpose	(1) Verify correct fit for purpose assessment based on eligibility criteria of study (2) Verify correct run of programs for fit for purpose (3) Verify selection of subset of matched patients	(1) Define selection / eligibility criteria (1) Selection / eligibility criteria reviewed and approved by content knowledge experts (2) Follow good programming guidelines (2) Validate against fit-for purpose specifications (2) Verify correct run of fit for purpose programs (2) Notify when validation is ready and programs are both signed off by programmer and QC (3) Align output with content knowledge experts
Transformation & mapping policy 2 (OMOP-SDTM)	(1) Mapping input created as input (2) Main programmer creates mapping programs from FHIR to OMOP+ (3) QC procedure	(1) Mapping files created directly linking from anonymized OMOP+ content to SDTM (1) Mapping files transferred to version controlled GitHub (2) Follow good programming guidelines (2) Input of latest version of mapping files and indicated SDTM CT version (2) Validate against mapping specifications (2) Verify correct run of OMOP+ to SDTM programs (2) Notify when programs are ready (3) Verify alignment with SDTM CT concept library (3) Verify all selected patients coming out of the fit for purpose assessment are included (3) Verify correct run of OMOP+ to SDTM programs (3) Verify no data is changed by mapping backwards (3) Verify that programs are both signed off by programmer and QC

LINEAGE

Based on these and the corresponding input, programs and output, the following lineage graph was created:
This graph shows the lineage between the different policy steps (yellow nodes), documents, mapping files, programs, and outputs. The graph is created by the governance tool from EQTY Life Sciences which links and controls the whole process enabling traceability and verifiability using modern but well-recognized techniques like cryptographic certifications.
Like fingerprints, cryptographic certifications generate unique derivatives systematically, making any data point generated by digital health technologies (DHT), applications, or systems a trusted source of truth. Weaving these encrypted and immutable certificates together, a verified data layer is created over entire RWE/AI workflows.



Certifying data, transformations and controlling data utilization rights at the variable level based on flexible patient consent, from capture through submission, sponsors, CROs and data providers now have the tools to prove to reviewers they acted responsibly. Providing auditors read-only rights to a sponsor's cryptographically verified data mesh, reviewers have the capability to seamlessly backward data map, click into data tables, trace data and processes backward to the source.

BACKWARDS MAPPING

The linkage from FHIR source to the final SDTM datasets and the corresponding mapping files ensure that the data was mapped backward to the original FHIR datasets. Based on the mapping files and de-anonymization, most of the original FHIR message could be recreated. The FHIR JSON code snippet below shows what could be reconstructed in black.


```
"resource": {
      "resourceType": "Observation",
       "id": "b02036b7-943d-abea-16f3-ac146af2eddb",
       "meta": {
       "profile": [ "http://hl7.org/fhir/us/core/StructureDefinition/us-core-bmi" ]
       },
      "status": "final",
       "category": [ {
         "coding": [ {
           "system": "http://terminology.hl7.org/CodeSystem/observation-category",
           "code": "vital-signs",
           "display": "Vital signs"
         } ]
       } ],
      "code": {
         "coding": [ {
           "system": "http://loinc.org",
       "code": "39156-5",
           "display": "Body mass index (BMI) [Ratio]"
         } ],
         "text": "Body mass index (BMI) [Ratio]"
      },
       "subject": {
       "reference": "urn:uuid:ad4c1e30-ddd7-5a4b-6af1-efcb19d8d96d"
```

```
      },
      "encounter": {
        "reference": "urn:uuid:110a54bd-0ef2-1d33-9a4d-acf6cd109be1"
      },
      "effectiveDateTime": "2021-08-23T08:19:34+01:00",
      "issued": "2021-08-23T08:19:34.512+01:00",
        "valueQuantity": {
        "value": 29.36,
        "unit": "kg/m2",
        "system": "http://unitsofmeasure.org",
        "code": "kg/m2"
      }
    },
    "request": {
      "method": "POST",
      "url": "Observation"
    }
}
```

Information that was not reconstructed (in red) was not mapped in the first place because of proportionality or confidentiality reasons or was deemed superfluous.

CONCLUSION

The Study design, stored in USDM format is able to drive the downstream processes like automated patient selection from the source, eligibility evaluation, visit selection and SDTM creation.
Cryptography serves as the foundation of modern audit tools by providing regulators with the highest level of confidence that real-world evidence has been responsibly captured, standardized, anonymized, and analyzed without bias. When data providers, vendors, and sponsors adopt open-source verification methods, they demonstrate control over their data and processes, offering regulatory reviewers verifiable proof of data quality and integrity.

REFERENCES

1. HMA Heads of Medicines Agencies, European Medicines Agency. Data quality framework for medicines regulation. [Online] https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-quality-framework-medicines-regulation. EMA/326985/2023.

2. Food and Drug Administration CDER, CBER and OCE. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory DEcision-Making for Drug and Biological Products. [Online] 07 2024. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory. FDA-2020-D-2307.
3. Synthea TM. [Online] https://syntheticheath.github.io/synthea/.
4. OMOP common data model v5.4. [Online] https://ohdsi.github.io/CommonDataModel/cdm54.html.
5. Digital Data Flow. [Online] https://www.cdisc.org/ddf.

CONTACT INFORMATION

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

Author Name: Berber Snoeijer
Company: ClinLine
Address: Van der Valk Boumanweg 15
Work Phone: +31 6 21197621
Email: b.snoeijer@clinline.eu
Website: www.clinline.org

Author Name: Alistair Dootson
Company: EQTY Life Sciences
Address: Innovate Carolina Junction- 136 E Rosemary St, Chapel Hill, NC 27514:
Email alistair.dootson@eqtylab.io:
Website:www.eqtylab.io