Paper RE07

Incorporating Real-World Data in a Low Interventional Study: Challenges and Considerations

Sandra VanPelt Nguyen, Pfizer, Inc.

ABSTRACT

This presentation will share an example of a low interventional study that incorporates real-world data along with conventionally collected clinical data to demonstrate real-world evidence of vaccine efficacy. We will explain the use of tokenization to link RWD from external data sources with the clinical database while maintaining patient privacy, how the data was mapped to SDTM, and challenges and considerations for mapping and analyzing the data.

INTRODUCTION

This paper is intended to share experience from a Phase 4 observational study for post-approval evaluation of the effectiveness of a vaccine in which a real-world data (RWD) component was incorporated. The study in this example is a "low interventional" study in which only minimal study procedures are performed. There is no vaccine administered to participants as a study procedure; instead, vaccination history is recorded for each enrolled participant.

Obtaining a complete, documented vaccination history for each participant is critical to success of the study. Sites are selected based on their experience and ability to obtain vaccination history data from various sources, however it is likely that some portion of participants will not have documented vaccination data available. It is also possible that prior vaccination may not be documented within the records that are available to the sites. Inclusion of individuals in the analysis who appear to be unvaccinated or have misclassified vaccine history could underestimate vaccine efficacy (VE), so any participants without fully documented vaccination history will be excluded from analysis. Therefore, accurate and complete prior exposure measurement requires all potential sources of vaccination history to be explored.

Due to this need, the study team decided to work with a RWD vendor to obtain vaccination history data from commercial data sources to supplement data obtained by the investigative sites. A tokenization process was utilized to enable privacy-protecting data linkage from the clinical study to tokenized RWD. The combined vaccination history data was mapped into SDTM and ADaM for use in analysis/reporting and regulatory submission. As RWD structures are not well-aligned to CDISC standards, this presented various challenges to address which are described below.

STUDY OVERVIEW

STUDY DESIGN

This study has a prospective test-negative design (TND), in which participants are enrolled based on a clinical case definition (relevant signs and symptoms of disease), with laboratory testing subsequently used to distinguish which patients were cases and which were controls. Test-negative controls are, by definition, derived from participants presenting with similar clinical syndromes as to the cases and differ only on the specific pathogen, serotype, or strain identified. The study is conducted in a real world setting because there are similar vaccines available in the market and it would be unethical and infeasible to conduct a randomized clinical trial.

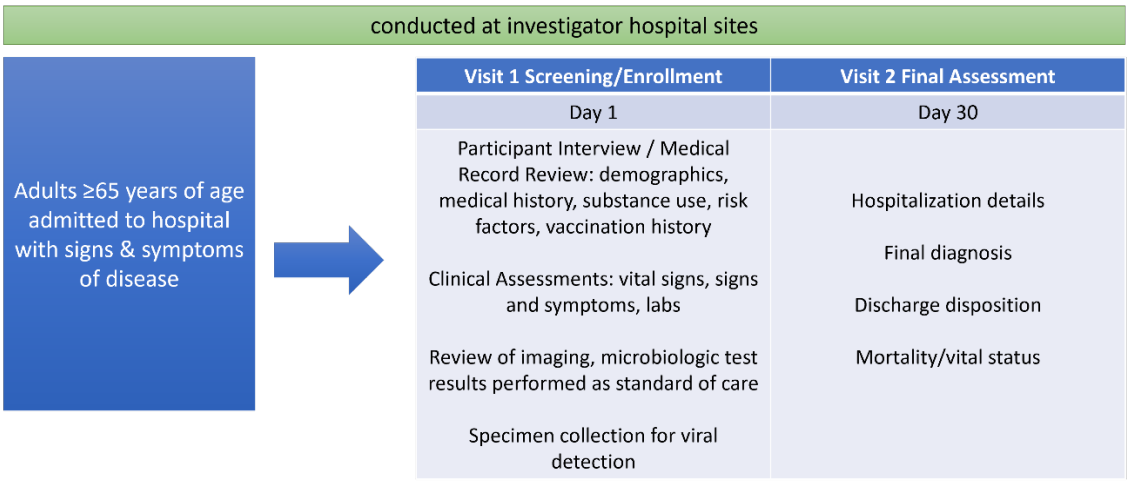
Like all observational designs, the TND still requires assessment for bias and confounding that may exist in the absence of randomized participation and blinded follow-up. Hence, this study is collecting data for each participant to assess and control for potential confounding factors, such as age and underlying conditions. As it is expected that persons at higher risk of disease are more likely to be vaccinated (potentially reducing the measured VE), VE estimates will be adjusted for confounding underlying conditions. By enrolling participants presenting with the same clinical syndrome, TND helps avoid bias due to unmeasured healthcare-seeking behavior, and the control group should be representative of the source population, reducing potential selection bias.

The study is being conducted at hospital sites throughout the United States that have past experience in similar studies and access to vaccination history data. Study participants will be adults at least 65 years of age who have been hospitalized at one of the sites with physician clinical suspicion of disease and at least two of the pre-specified clinical symptoms.

STUDY ASSESSMENTS

Visit 1 will occur within 1-2 days post-admission. Apart from urine specimen collection for virus detection, all other clinical and medical data collection is done through participant interview and review of medical records. Participants are asked to provide vaccination history information as well as contact information for primary care physician(s), pharmacy, and health insurance providers so that site staff can obtain documented vaccination history data. State registries may also be explored, if available.

Vaccination records review should be completed by the end of the Day 30 visit window, which should allow time for sites to check the available sources. Data obtained by the sites will be entered into Concomitant Medications case report forms (CRFs) in the study database.



Success of the study depends largely on adequate uptake and use of the vaccine in adults ≥65 years of age, accrual of cases needed to evaluate the primary endpoint, and identification of documented vaccination history. Participants will be considered vaccinated if the vaccine is documented as being received >30 days before hospitalization and within the last five years. To improve exposure ascertainment and minimize exposure measurement error, Pfizer decided to supplement site-collected vaccine history data with commercially available tokenized RWD from one or more commercial electronic health record (EHR) data providers.

TOKENIZATION AND REAL-WORLD DATA FLOW

WHY USE TOKENIZATION?

Most Americans receiving health care services now have their health data recorded electronically, and commercial solutions have emerged to enable participant-level linkage to commercially available de-identified data assets. One such solution utilizes tokenization which creates a persistent, linkable patient identification without the need for direct identifiers. Through tokenization, identifiable patient information is transformed into an irreversible random string of characters that has no meaningful value on its own but can be used to link patient-level data across different datasets.

Tokenization provides a secure and consistent mechanism for linking data without compromising the privacy of study participants, because the tokens are:

- irreversible and blinded: input personally identifiable information (or PII) cannot be backward engineered from the output token.
- unique: each token is unique to the input PII, avoiding false matches.
- consistent: the same set of input PII will always create the same tokens.
- secure: participant PII never leaves the study site.

This study is piloting the use of tokenization and will conduct supplemental analyses to demonstrate the success rate for this data.

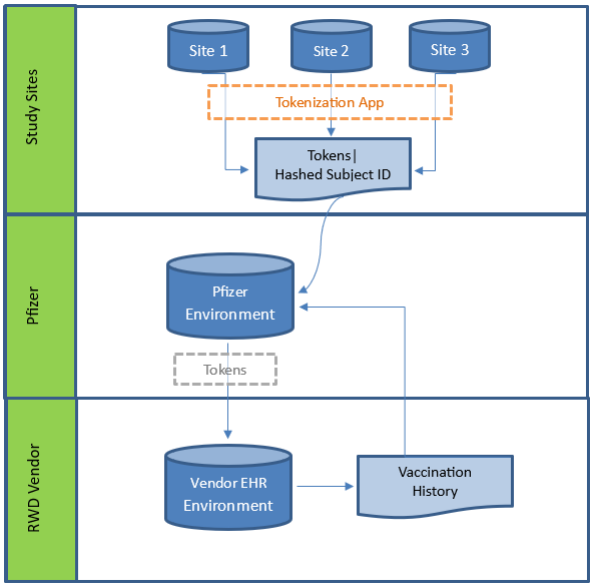
TOKENIZATION PROCESS

Participants can opt in or out for tokenization when completing informed consent. Site staff will then generate tokens by accessing the token generation application made available by the vendor. Each individual element of PII is an individual token and multiple tokens can be created from different PII elements, improving the precision and granularity of matching. While PII such as Social Security Number (SSN) can create very accurate matches, it is often missing in RWD sources because it is a more sensitive identifier that people do not want to share, thus it could actually limit matches. Hence PII used for this study include: First Name, Last Name, Date of Birth (DOB), Biological Sex and Zip Code.

On the RWD vendor's side, the technology de-identifies PII, replacing it with one or more tokens. Tokens are assigned based on the PII in the record so that the same tokens can be consistently created from any data set where the underlying PII is the same. Matching tokens can then be used to link a patient's records across data sets without ever exposing the PII of that patient. Multiple tokens are often created and used to improve matching.

DATA FLOW

Tokenization data is shared with Pfizer and provided to the RWD vendor who then uses it to gather vaccine history data based on Pfizer's specifications and transmits the data to Pfizer for use in the study. This is managed as non-CRF data that goes through the external data transfer process. The tokens are then used to connect the RWD with the participant data in the study database. Any discrepancies between the RWD and the site-collected data are reconciled through queries to the site. Because the RWD is considered supplemental to site-collected data, the site-collected data is prioritized over tokenized data when there are unresolved data discrepancies.



The participant-specific tokenized RWD is managed separately from site-collected data and use of the data in reporting is dependent on being able to successfully link tokenized RWD with study participant data.

MAPPING AND ANALYSIS CONSIDERATIONS

DATA STANDARDS

FDA guidance[1] on data standards for submissions containing RWD indicate that the data is subject to the standards defined in the Data Standards Catalog[2], meaning SDTM and ADaM standards should be used unless a waiver has been granted. FDA does recognize that there are challenges involved in standardizing study data derived from RWD sources, including de-identification methodologies used to protect patient data when shared. It is important that sponsors document any changes to data to conform to the supported data standards, and the potential impacts of those changes. FDA notes that "with adequate documentation of the conformance methods used and their rationale, study data derived from RWD can be transformed into SDTM and ADaM datasets and submitted to FDA in an applicable submission" and recommends that sponsors discuss transformation and mapping plans for RWD with them in advance of a submission.

SDTM MAPPING CONSIDERATIONS

While it would be optimal for the data transfer to contain one record for each vaccination, we went with a vertical file structure that allowed for a variable number of tokens to be used (in case more than one RWD vendor was utilized). This resulted in multiple records being transferred for a vaccine (one for each token used in the matching process), which then need to be collapsed appropriately. Additionally, the same vaccine could be found by the RWD vendor in multiple commercial databases which may be using different medical coding systems, so we had to use caution when collapsing the data to maintain and distinguish each unique vaccine match in the SDTM dataset.

Besides the tokens, additional data fields were included which are not traditionally collected in a clinical trial, such as query date, index date, source type, code type and code, and "found" flags. Each of these data points had to be

assessed to determine whether they were needed in SDTM and where and how best to map them, in a way that would support analysis plans while complying with SDTM standards.

A list of vaccines and protocol-specified timeframe was provided to the RWD vendor. If a particular vaccine was not found, then a row of data was to be transferred to Pfizer indicating that no data was found. A combination of the "found" flag and "reason not found" (provided by vendor) had to be used to determine if the vaccine was not found at all, was found but excluded due to "collision" or "split" (e.g. if there are multiple people with the same name and DOB, where a "collision" indicates that one participant matched to multiple patients in the RWD and a "split" indicates that multiple patients in the RWD matched to one participant), or was found but not within the defined time period.

Since historical vaccine data is being collected (no vaccine administered during the study), Pfizer is mapping both site-collected and RWD to the CM domain in SDTM. It was also important to be able to distinguish between site-collected and RWD, as various rules need to be applied for handling duplicates/discrepancies for analysis and reporting.

Example of RWD (note: a subset of fields is shown below):

Row #	HASHED_SUBJECT_ID	PRODUCT_NAME	ADMIN_DT	CODE_TYPE	CODE
1	8ir5wljBi3Xc9/qoGWARXilAl605KE3esC+jBZRfXcM=	Vaccine ABC	2023-01-01	NDC	49281012265
2	8ir5wljBi3Xc9/qoGWARXilAl605KE3esC+jBZRfXcM=	Vaccine ABC	2023-01-01	NDC	49281012265
3	8ir5wljBi3Xc9/qoGWARXilAl605KE3esC+jBZRfXcM=	Vaccine ABC	2023-01-01	CPT	90732
4	8ir5wljBi3Xc9/qoGWARXilAl605KE3esC+jBZRfXcM=	Vaccine ABC	2023-01-01	CPT	90732
5	8ir5wljBi3Xc9/qoGWARXilAl605KE3esC+jBZRfXcM=	Vaccine XYZ		N/A	N/A
6	8ir5wljBi3Xc9/qoGWARXilAl605KE3esC+jBZRfXcM=	Vaccine XYZ		N/A	N/A
7	jRyaIIi9NjCt1aVQZ+z/HrggBlvNke1iFm8Ge2FdZpA=	Vaccine ABC		N/A	N/A
8	jRyaIIi9NjCt1aVQZ+z/HrggBlvNke1iFm8Ge2FdZpA=	Vaccine ABC		N/A	N/A
9	jRyaIIi9NjCt1aVQZ+z/HrggBlvNke1iFm8Ge2FdZpA=	Vaccine XYZ		N/A	N/A
10	jRyaIIi9NjCt1aVQZ+z/HrggBlvNke1iFm8Ge2FdZpA=	Vaccine XYZ		N/A	N/A
11	H9DtCKRiQeeIF/bZBbpto1oMhXm7uXW1z73O3OZu6DA=	Vaccine XYZ	2023-10-12	CPT	90732
12	H9DtCKRiQeeIF/bZBbpto1oMhXm7uXW1z73O3OZu6DA=	Vaccine XYZ	2023-10-12	CPT	90732

Row #	TKN_NUM	TKN_VALUE	FOUND_FLAG	NF_REASON
1	TOKEN1	ZMFfB8B7GFqEqNksw8rsrIuPCAWPBjmYxtUNbyTWGT0=	Y	
2	TOKEN2	LPhxlCrJuURlDKiJeR6JmdAde+RMA8JQ9RmRc4HEFCE=	Y	
3	TOKEN1	ZMFfB8B7GFqEqNksw8rsrIuPCAWPBjmYxtUNbyTWGT0=	Y	
4	TOKEN2	LPhxlCrJuURlDKiJeR6JmdAde+RMA8JQ9RmRc4HEFCE=	Y	
5	TOKEN1	ZMFfB8B7GFqEqNksw8rsrIuPCAWPBjmYxtUNbyTWGT0=	N	Token not found
6	TOKEN2	LPhxlCrJuURlDKiJeR6JmdAde+RMA8JQ9RmRc4HEFCE=	N	Token not found
7	TOKEN1	01u5ezuRo1y9/Y+wxRKE6ZVwgq3OjBsWtJ6EmZm/gRo==	N	Data excluded due to collision
8	TOKEN2	vuZKYQqOBw+KMc1kTCt4vTh3x8LkNi2LgOF96ep0xzM=	N	Data excluded due to collision
9	TOKEN1	01u5ezuRo1y9/Y+wxRKE6ZVwgq3OjBsWtJ6EmZm/gRo==	N	Token was found but the specified data was not present within the requested time period
10	TOKEN2	vuZKYQqOBw+KMc1kTCt4vTh3x8LkNi2LgOF96ep0xzM=	N	Token was found but the specified data was not present within the requested time period
11	TOKEN1	yZlPqWd/5oDpIwxrtpNltP2OEjKJ9QpAuV1btRU1+ns=	Y	
12	TOKEN2	k6XOA2uJvXgVXamhRjyk1oX0WT7XT9kBLQDgfe6XH/g=	Y	

Each highlighted pair of records represents one match (or lack of match), based on a set of tokens and code (used to search for the vaccine of interest), with each token in a separate row. If another token was created, additional rows could be added in the transfer without changing the file structure.

SDTM MAPPING EXAMPLE

This example combines the RWD above with CRF data for the same participants.

- Site-collected data is represented in rows 1-3, 7-8, and 11. The vaccines were pre-specified so the CMPRESP flag is set to "Y" and occurrence (CMOCCUR) is indicated based on what data the site was able to obtain.
    - Vaccine ABC was received twice within the last five years for Subject 1001 (rows 1 & 2). The 2023 vaccination was reported in the participant interview and the 2024 vaccination was proven with pharmacy records. Vaccine XYZ was not identified in any sources (row 3). Non-standard variable (NSV) CMSOINF (Source of Information, collected in the CRF) in SUPPCM indicates the source used by the site to verify vaccination.
    - For Subject 1002, Vaccine ABC and XYZ (rows 7 & 8) were not reported by the participant or found during records review.
    - Vaccine XYZ was not found for Subject 1003 (row 11).

- From the RWD, 12 records in the data transfer were collapsed into six records in SDTM (rows 4-6, 9-10, and 12). The NSV RWDFL (Real-World Data Flag) is used to flag the RWD records in SUPPCM.
  - For Subject 1001, Vaccine ABC was found twice based on different medical coding systems (rows 4 & 5). Pfizer decided to maintain these as two separate records in the CM domain since information could potentially differ across different sources. The NSV TKNCDTYP (Code Type) indicates the medical coding system that was used: NDC = National Drug Code Directory and CPT = Current Procedural Terminology. The NDC and CPT codes then had to be mapped to WHO Drug.
  - Vaccine XYZ for Subject 1001 and Vaccine ABC for Subject 1002 were not found by the RWD vendor as indicated by the found flags and reasons not found provided in the data transfer (rows 6 & 9). Since these values were pre-specified and the data was needed for analysis, Pfizer decided to map any "not found" vaccines using CMSTAT = "NOT DONE" rather than CMOCCUR = "N" as it could not be assumed that the vaccine had not been received just because it was not found by the vendor or had collisions or splits.
  - Vaccine XYZ for Subject 1002 (row 10) was found by the RWD vendor but was not received within the last five years.
  - Vaccine XYZ for Subject 1003 (row 12) was found by the RWD vendor.

cm.xpt:

Row	STUDYID	DOMAIN	USUBJID	CMTRT	CMDECOD	CMCAT	CMPRESP	CMOCCUR
1	ABC001	CM	1111 1001	ABCVAX	Vaccine ABC	VACCINATIONS	Y	Y
2	ABC001	CM	1111 1001	ABCVAX	Vaccine ABC	VACCINATIONS	Y	Y
3	ABC001	CM	1111 1001	XYZVAX	Vaccine XYZ	VACCINATIONS	Y	N
4	ABC001	CM	1111 1001	ABCVAX	Vaccine ABC	VACCINATIONS	Y	Y
5	ABC001	CM	1111 1001	ABCVAX	Vaccine ABC	VACCINATIONS	Y	Y
6	ABC001	CM	1111 1001	XYZVAX	Vaccine XYZ	VACCINATIONS	Y	
7	ABC001	CM	1111 1002	ABCVAX	Vaccine ABC	VACCINATIONS	Y	N
8	ABC001	CM	1111 1002	XYZVAX	Vaccine XYZ	VACCINATIONS	Y	N
9	ABC001	CM	1111 1002	ABCVAX	Vaccine ABC	VACCINATIONS	Y	
10	ABC001	CM	1111 1002	XYZVAX	Vaccine XYZ	VACCINATIONS	Y	N
11	ABC001	CM	1111 1003	XYZVAX	Vaccine XYZ	VACCINATIONS	Y	N
12	ABC001	CM	1111 1003	XYZVAX	Vaccine XYZ	VACCINATIONS	Y	Y

Row	CMSTAT	CMREASND	CMSTDTC	CMEVLINT	SUPPCM. CMSOINF	SUPPCM. RWDFL	SUPPCM. TKNCDTYP
1			2023-01	-P5Y	SELF-REPORT		
2			2024-01-07	-P5Y	PHARMACY RECORD		
3				-P5Y			
4			2023-01-02	-P5Y		Y	NDC
5			2023-01-02	-P5Y		Y	CPT
6	NOT DONE	Token not found		-P5Y		Y	
7				-P5Y			
8				-P5Y			
9	NOT DONE	Data excluded due to collision		-P5Y		Y	
10				-P5Y		Y	
11				-P5Y			
12			2023-10-12	-P5Y		Y	CPT

The use of CMOCCUR and CMSTAT here are not optimal, but seemed to be the best fit using the available standard variables.

ANALYSIS CONSIDERATIONS

Vaccine effectiveness is assessed by comparing the number of confirmed cases of disease in the vaccinated group against the unvaccinated group. Both crude and adjusted (for confounding factors) VE will be generated. Test-positive cases include participants diagnosed with the disease in which the virus has been detected using a urinary antigen detection (UAD) assay performed on the urine specimen. Any specimens which do not detect the virus or have indeterminate outcome are included in the control group.

Analysis Populations:

- All Enrolled
- Evaluable (enrolled + documented vaccination history + UAD test performed)

Prior vaccination is summarized as a binary (Yes/No) variable with participants categorized as "ever vaccinated" or "never vaccinated" for each vaccine. All available vaccine history data must be assessed to determine if the vaccine has been received. Because tokenized data are considered supplemental to site-collected data, these data are prioritized over tokenized data when there are unresolved data discrepancies. In cases where multiple vaccine records are available, the following rules are applied:

- If site-reported data includes vaccine(s) of interest, but tokenized data does not, then site-reported data will be used.
- If site-reported data does not include the vaccine(s) of interest, but tokenized data does, then tokenized data will be used.
- If site-reported data and tokenized data both include vaccine(s) of interest and vaccine type and date match, then site-reported data will be used.
- If site-reported data and tokenized data both include vaccine(s) of interest, but either vaccine type or date do not match and remain discrepant after all queries for vaccine type and date are resolved, then site-reported data will be used.
- If neither site-reported data nor tokenized data sets include evidence of vaccine(s) of interest and site-reported data search is complete, then participant will be categorized as having no evidence of vaccination.

Furthermore, any records with a collision or split are considered un-linkable.

In addition to basic medications summaries and the vaccine effectiveness analysis, the following summaries will be generated to assess the use of tokenization:

- Percentage of sites and participants consented to the tokenization process
- Percentage of tokenized participants that are linked to a record from commercially available RWD sources, along with collision and split rates
- Percentage of participants for whom prior vaccine exposure is identified by the tokenized RWD but not by site, and vice versa
- Percentage of participant cases with discordant site-reported and tokenized RWD vaccine history data and cases where there is alignment.

ANALYSIS DATA EXAMPLE

This example uses the SDTM data from above, however data is sorted differently to group each subject/vaccine together and select the records to be used in analysis. Efficacy will be analyzed for "any" prior vaccination and for each specific vaccine. Analysis flags (ANLzzFL) are used to flag the records that will be used in analysis:

- ANL01FL: flags the first of any vaccination received (or, if there are no records with CMOCCUR = "Y", the record indicating that vaccine was not received or status was indeterminate) based on the prioritization defined above. This is flagged once per subject.
- ANL02FL: similar to ANL01FL, but flags the first occurrence of Vaccine ABC that meets the criteria to be included in analysis. This is flagged once per subject per vaccination type.
- ANL03FL: similar to ANL02FL, but for Vaccine XYZ.

For row 1, there was a partial vaccination date in SDTM, so a complete date is imputed in ASTDT. In practice, we would expect this discrepancy (there was a full data available in the RWD) to be queried, with a correction by the site. According to the decision rules, the site-reported record takes priority.

Rows 11 and 12 provide an example of where the tokenized data is used over the site-reported data.

adcm.xpt:

Row #	STUDYID	USUBJID	SUBJID	SITEID	CMTRT	CMDECOD	CMPRESP	CMOCCUR	CMSTAT	CMREASND
1	ABC001	1111 1001	1001	1111	ABCVAX	Vaccine ABC	Y	Y		
2	ABC001	1111 1001	1001	1111	ABCVAX	Vaccine ABC	Y	Y		
3	ABC001	1111 1001	1001	1111	ABCVAX	Vaccine ABC	Y	Y		
4	ABC001	1111 1001	1001	1111	ABCVAX	Vaccine ABC	Y	Y		
5	ABC001	1111 1001	1001	1111	XYZVAX	Vaccine XYZ	Y	N		
6	ABC001	1111 1001	1001	1111	XYZVAX	Vaccine XYZ	Y		NOT DONE	Token not found
7	ABC001	1111 1002	1002	1111	ABCVAX	Vaccine ABC	Y	N		
8	ABC001	1111 1002	1002	1111	ABCVAX	Vaccine ABC	Y		NOT DONE	Data excluded due to collision
9	ABC001	1111 1002	1002	1111	ABCVAX	Vaccine XYZ	Y	N		
10	ABC001	1111 1002	1002	1111	ABCVAX	Vaccine XYZ	Y	N		
11	ABC001	1111 1003	1003	1111	ABCVAX	Vaccine XYZ	Y	N		
12	ABC001	1111 1003	1003	1111	ABCVAX	Vaccine XYZ	Y	Y		

Row #	CMSTDTC	ASTDT	ASTDTF	RWDFL	TKNCDTYP	ANL01FL	ANL02FL	ANL03FL
1	2023-01	01JAN2023	D			Y	Y	
2	2024-01-07	07JAN2024						
3	2023-01-02	02JAN2023		Y	NDC			
4	2023-01-02	02JAN2023		Y	CPT			
5								Y
6				Y				
7						Y	Y	
8				Y				
9								Y
10				Y				
11								
12	2023-10-12	12OCT2023		Y	CPT	Y		Y

In ADSL, flags are assigned to inform whether the participant is included in the Enrolled analysis population (ENRLFL) and the Evaluable population (EVALFL). Vaccination flags ANYFXFL (any vaccination), VAX1FL (Vaccine ABC), and VAX2FL (Vaccine XYZ) indicate whether the participant was vaccinated or not, according to the available data and decision rules. In cases where vaccine history is not available, the vaccination flags are null.

- For Subject 1001, the site reported that Vaccine ABC had been received within the last five years, but not Vaccine XYZ.
- For Subject 1002, vaccination records were searched but no acceptable matches were found by the site or in the RWD, so the participant is evaluable for all vaccination analyses, but considered unvaccinated for each type.
- Subject 1003 only had data provided for Vaccine XYZ. The RWD found a match, so the participant is evaluable for Vaccine XYZ, but is not evaluable for Vaccine ABC since no documented history was recorded.

adsl.xpt:

STUDYID	USUBJID	SUBJID	SITEID	ENRLFL	EVALFL	ANYVXFL	VAX1FL	VAX2FL
ABC001	1111 1001	1001	1111	Y	Y	Y	Y	N
ABC001	1111 1002	1002	1111	Y	Y	N	N	N
ABC001	1111 1003	1003	1111	Y	Y	Y		Y

CONCLUSION

The incorporation of RWD can be extremely helpful for identifying the complete picture for data components of interest in a clinical trial, and can potentially reduce the burden on participants and sites. The use of tokenization can ease privacy concerns and reduce hurdles to obtaining and integrating commercial data. We are hopeful that inclusion of RWD will also improve quality and completeness of the data. Despite these benefits, there remain many challenges in terms of mapping the data to accepted data standards and analyzing the data. This requires a clear understanding of each data field in the RWD and how the data needs to be used for analysis and reporting, plus taking necessary steps to "ensure the reliability and relevance of the data"[3]. Clear and detailed documentation for is needed for each step in the data flow along with careful assessment and planning. Pfizer's pilot demonstrates that RWD can successfully be mapped to CDISC data standards for reporting and submission.

REFERENCES

[1]US Department of Health and Human Services, Food and Drug Administration. *Data Standards for Drug and Biological Product Submissions Containing Real-World Data: Guidance for Industry.* December 2023. Available at: https://www.fda.gov/media/153341/download

[2]US Department of Health and Human Services, Food and Drug Administration. *Data Standards Catalog*. Available at: https://www.fda.gov/media/160564/download

[3]US Department of Health and Human Services, Food and Drug Administration. *Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory DecisionMaking for Drug and Biological Products: Guidance for Industry* (Draft). September 2021. Available at: https://www.fda.gov/media/152503/download

ACKNOWLEDGMENTS

Thank you to Pam Singletary, Mike Hamidi, and June Cheng for helping me understand the tokenization process and data flow, providing clarifications, and/or providing supporting materials for this paper. I also thank Yogesh Gupta for reviewing and providing feedback.

RECOMMENDED READING

This paper is an expansion on a presentation that was part of the PHUSE Real World Evidence Community Forum "Experiences with Adapting RWD to CDISC Submission Standards" on October 5, 2023. The slides and recording are available at: https://advance.phuse.global/pages/viewpage.action?pageId=80806198

To learn more about how tokenization can be used to match de-identified data across different data sets, I recommend reading Datavant's white paper "*Overview of Datavant's De-Identification and Linking Technology for Structured Data"* which is available at: https://datavant.com/wp-content/uploads/dlm_uploads/2018/09/WhitePaper_-De-Identifying-and-Linking-Structured-Data.pdf

CONTACT INFORMATION

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

Author Name: Sandra VanPelt Nguyen
Company: Pfizer, Inc.
Email: sandra.vanpeltnguyen@pfizer.com