The Curious Case of Observational Studies - A Programmer's Dilemma

Vikas Shokeen, Novartis, Hyderabad, India

ABSTRACT

There are two major types of studies in clinical research: Observational Studies (OS) and Randomized Clinical Trials (RCT). While the Randomized trials have been titled as the 'Gold Standard' in terms of producing reliable outcomes, Observational studies have also proved beneficial in various cases especially, when there is little to no evidence of the mechanism of a disease. However, these studies have put the statistical programmers in a state of conundrum due to its various challenges such as 'No pre-defined CDISC Data Standards' etc. Despite these obstacles, programmers navigate through this turmoil with utmost care and following the CDISC guidelines of RCT's. This paper will outline the programming deterrents of observational studies, highlight the key takeaways with the help of an example of a non-experimental cohort study and suggest how methodology studies could be enhanced to pave way for a smoother clinical journey.

INTRODUCTION

Randomized Clinical trials represent the gold standard in research designs and exert a significant influence on medical guidelines and subsequent clinical practices. They adhere to well-established CDISC standards and predefined submission guidelines. In contrast, "Observational studies" lack clearly defined standards for data collection and analysis. This creates challenges for statistical programmers, leading to reduced confidence in the analysis results.

Observational studies encompass various types, including Case-Control, Cohort, Cross Sectional, and Ecological, which are further categorized into retrospective and prospective studies. Among these, cohort-defined prospective observational studies are generally less burdensome for statisticians and statistical programmers due to their defined covariates and baseline outcomes. However, the absence of defined CDISC standards and the ambiguity in preparation of clinical documents like Statistical Analysis Plans (SAP) pose challenges for programmers in developing Study Data Tabulation Models (SDTM), Analysis Dataset Models (ADaM), and their corresponding outputs.

This paper aims to highlight the hurdles faced by programmers in a cohort-defined prospective observational study, as well as identify issues encountered in clinical documents such as SAP (Statistical Analysis Plan) and TFL (Tables, Figures, and Listings) Shells. Additionally, the paper will offer potential solutions to these challenges.

Before we proceed with the detailed examination of a Cohort Defined Prospective Observational Study, it is imperative to acknowledge that Cohort-based studies hold a superior position, particularly in exploratory research. Additionally, a comprehensive understanding of the differentiation between prospective and retrospective studies is essential.

RATIONALE BEHIND COHORT STUDIES

Cohort-defined observational studies are a type of research design used in epidemiology to investigate the relationships between specific exposures and subsequent outcomes. They involve identifying a group of individuals, known as a cohort, who share a common characteristic or have been exposed to a particular factor of interest. This cohort is then followed over an extended period, often spanning years or even decades, to observe and record various outcomes. These outcomes may include the incidence of diseases, changes in health status, or other relevant events.

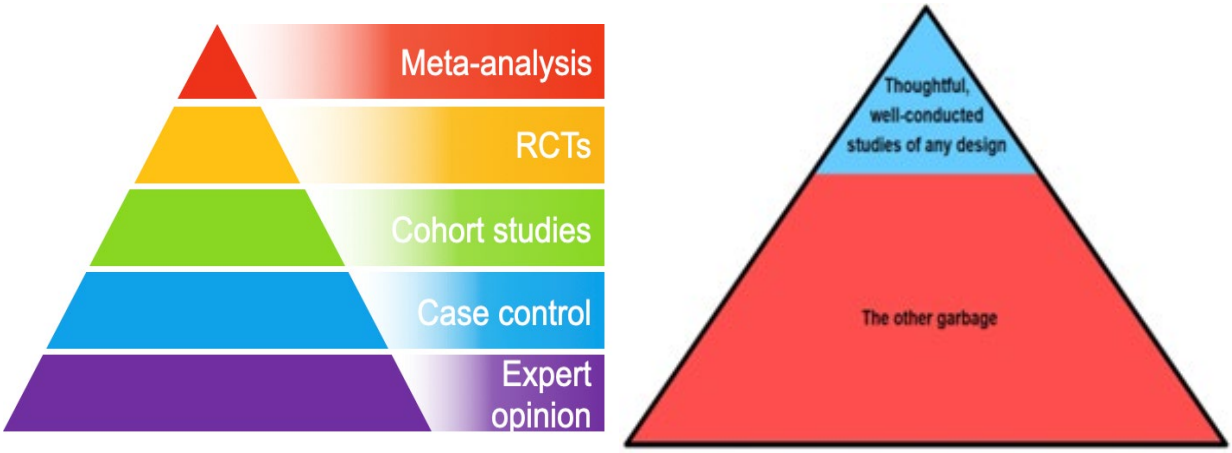
The key characteristic of cohort studies is their longitudinal nature, meaning that data is collected over time. Researchers focus on assessing exposure to a specific variable or risk factor, such as occupation, age, lifestyle, or

exposure to a particular substance. Additionally, covariate information is often collected to account for potential confounding variables that may influence the relationship between the exposure and outcome.

Cohort studies can be categorized as either prospective, where the study begins in the present and follows individuals into the future, or retrospective, where historical data is used to assemble a cohort, and outcomes are then examined backward in time. They offer distinct advantages, including the ability to study exposures that are difficult or unethical to randomize, the representation of the study population in relation to the target population, and the calculation of incidence rates.

Overall, cohort-defined observational studies are valuable tools for investigating causal relationships between exposures and outcomes, particularly for rare exposures or events, and for assessing the long-term effects of exposures on health. Examples of cohort studies include the Framingham Heart Study, occupational cohort studies, and birth cohort studies.

The pyramid illustrates the hierarchy of credibility that different types of studies hold in the clinical realm. At the pinnacle of observational research, cohort studies are positioned as the most esteemed.



Now, let's delve into the specifics of a Cohort Defined Prospective Observational Study.

RATIONALE AND OBJECTIVE

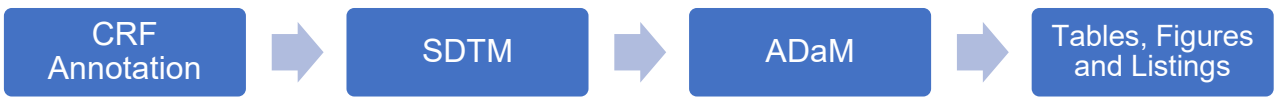
This natural history study will better characterize the progression of participants with a disease X through longitudinal study of morphologic ocular imaging (meibography), evaluation and understanding of clinical disease severity, and ocular surface disease endpoints across a diverse ethnic population. It will help in the selection of the appropriate population for future studies, appropriate endpoints, and insights to potential future therapeutic interventions.

- The primary objective is to evaluate the rate of disease progression over time based on meibography images.

STUDY DESIGN

The study will entail an observational period of up to three years and will encompass two distinct groups: individuals diagnosed with disease X and a control cohort comprising "normal" healthy volunteers, defined as those exhibiting a healthy ocular surface without any indications or symptoms of disease X. Furthermore, the study will involve stratification based on race, with participants classified into two racial groups. This stratification will be contingent upon the investigator's assessment of disease severity, which will be categorized as normal, mild, moderate, or severe. Given the absence of globally recognized definitions for the disease, the stratification process will rely on the investigator's professional judgment of disease severity.

Below is a visual representation of key milestones in the realm of Statistical Programming:



ANNOTATED CRF

An Annotated Case Report Form (aCRF) is a PDF document that maps the clinical data collection fields used to capture subject data (electronic or paper) to the corresponding variables or discrete variable values contained within the SDTM datasets.

It's not unusual for observational studies to encounter difficulties and discussions when mapping and organizing data, especially when striving to adhere to SDTM standards and optimize data for subsequent analysis. Let's break down the scenario and our proposed resolution:

Challenges Encountered

- ***Utilization of Decommissioned Domain (MO):*** The utilization of the MO (Morphology) domain, which has been retired since SDTMIG v3.3, can lead to ambiguity and debates. This domain might have limitations and may not fully align with the data requirements of the study.

- ***Redundancy in Test Data:*** When the same test data is recorded in both the MO (Morphology) and ZX (Examination Results) domains, it can result in duplication and complications during data analysis, potentially influencing data quality and interpretation.

Proposed Resolution

Custom Analysis Dataset: The creation of a custom Analysis dataset to amalgamate the pertinent test data from both the MO and ZX domains is a practical solution. This approach allowed for the consolidation of duplicated data and streamlines it for subsequent analysis.

Queries

- Was it possible to anticipate and avert this scenario?
  In hindsight, the avoidance of duplicating test data between the MO and ZX domains could have been attainable with meticulous planning and the design of the CRFs. Ensuring that data elements are captured within a single domain, or that there is a compelling justification for data to exist in both domains, can help preclude this issue.

- Were there other potential methods or strategies that could have been considered?
  Creating a custom Analysis dataset is a reasonable approach when confronted with data duplication between domains. Nevertheless, the best practice is to avert such duplication from occurring initially. Guaranteeing that CRFs are designed with precision and in adherence to SDTM standards is imperative.

- Is This Issue Prevalent in All Observational Studies?
  The specific challenges experienced in observational studies can fluctuate based on the study's configuration, data origins, and objectives. While some challenges, like data variability and non-standardization, are widespread in many observational studies, the problem of data duplication across domains may be exclusive to studies or circumstances.

Key Insights

- **Pre-Define CRF Elements**: Clearly define the data elements to be collected, including variables, their permissible values, units of measurement, and any associated validation rules. Use standardized terminology and coding systems where applicable.
- **Effective CRF design and mapping** are pivotal for the triumph of observational studies, as they exert an influence on data quality and subsequent analysis. When the CRF is well-annotated, information flows seamlessly and effectively into the SDTM.
- **Reference Therapeutic Area User Guides:** When SDTM standards lack clarity or guidance for specific domains, it can be beneficial to consult relevant Therapeutic Area User Guides (TAUGs). These guides provide domain-specific recommendations and examples for mapping variables, ensuring consistency with industry practices.
- **Utilize Existing SDTM Domains:** Instead of creating custom domains when SDTM definitions are unclear, explore the possibility of using existing SDTM domains that closely align with the data collected. In our example, using the OE (Ophthalmic Examination) domain might have been a suitable alternative for capturing the required data.
- **Custom Analysis datasets** can be a valuable resource for harmonizing data when duplication arises, yet it is preferable to prevent duplication through well-considered CRF design.

To conclude, Collaborative efforts among data managers, statistical programmers, and experts in the field are essential for addressing mapping complications and upholding data quality in observational studies.

STUDY DATA TABULATION MODEL (SDTM)

SDTM provides a standard for organizing and formatting data to streamline processes in collection, management, analysis, and reporting. Implementing SDTM supports data aggregation and warehousing; fosters mining and reuse; facilitates sharing; helps perform due diligence and other important data review activities; and improves the regulatory review and approval process. SDTM is also used in non-clinical data (SEND), medical devices and pharmacogenomics/genetics studies.

Statistical programmers face three kinds of challenges when creating the SDTM:

- Maintaining data quality is a significant consideration when working with the SDTM in observational studies where cohorts are defined.
- Ensuring that the SDTM structure adheres to CDISC standards is crucial. This is closely linked to the accuracy and completeness of the Annotated CRF.
- Verifying the presence of both mandatory and anticipated variables across different domains is essential.

Here are the SDTMs created for our study, and we will be going over specific ones in depth.

Intervention Domains	Procedure (PR)
	Concomittant Medications (CM)

If the CRF is annotated in accordance with established protocols, no substantial concerns arise with regards to these domains.

Special Purpose Domains	Demographics (DM)
	Subject Elements (SE)
	Subject Visits (SV)

The Demographics (DM) domain raised notable concerns, particularly pertaining to variables such as ARM/ARMCD, RFSTDTC, RFXSTDTC, and so forth. These will be addressed in detail at a later stage.

Findings Domains	Death (DD)
	Inclusion/Exclusion Criteria Not Met (IE)
	Questionnaire (QS)
	Morphology (MO)
	Disease Response (RS)
	Subject Status (SS)

The persistence of inconsistent data represents a paramount concern during the formulation of findings domains. For example, the handling of Questionnaire data (QS) demands substantial diligence from programmers to verify the comprehensive and accurate addressing of all study-related questions. Another noteworthy concern within the findings domains is the complexity associated with mapping.

Trial Design Domains	Trial Summary (TS)
	Trial Arm (TA)
	Trial Element (TE)
	Trial Visits (TV)
	Trial Inclusion/Exclusion Criteria (TI)

The inclusion of Trial Summary poses no issue. However, there are certain considerations to be addressed regarding the representation of variables in domains such as Trial Arm (TA). Additionally, executing Trial Design domains proves to be straightforward in cohort-defined prospective observational studies.

Custom Domains	Ophthalmic Examination Results (ZX)

The ZX domain was established to encompass all ophthalmic examination results. This, however, led to discrepancies in the data between the MO and ZX domains, as identical information was included in both domains, as recommended earlier.

Event Domains	Adverse Events (AE)
	Disposition (DS)
	Medical History (MH)

The Disposition domain poses a challenge in observational studies, as the milestones may not be adequately defined in the protocol, potentially resulting in irregularities in the data.

The subsequent points outline some of the pivotal challenges we encountered during the development of the respective SDTMs. The ensuing variables gave rise to an array of issues.

Demographics (DM)

VARIABLE	CHALLENGE	SOLUTION
RFSTDTC (Subject Reference Start Date/Time)	Typically, this corresponds to the date and time of the subject's initial exposure to the study drug. However, it's important to note that observational studies do not involve any treatment administration. Consequently, there arises a question of what should be recorded in this variable. While it is an anticipated variable, its value may be null. Nonetheless, it would not be prudent to leave it blank for all records.	Given that RFSTDTC serves as the reference point for calculating study days, it assumes the role of an anchor in the study timeline. It would be practical to assign either the date of informed consent or the date of randomization to this variable.
RFENDTC (Subject Reference End Date/Time)	In the context of observational studies, where no treatment is administered, determining what should be entered in RFENDTC (the date/time of the subject's last exposure to the study drug) becomes a pertinent question. Since there is no drug exposure in observational studies, RFENDTC may be appropriately designated as null or left blank for all records.	Given that RFENDTC must be filled for all randomized subjects, a practical approach would be to populate it with the End of Study Date time value. This ensures consistent and complete data across all subjects in the study.
RFXSTDTC (Date/Time of First Study Treatment)	This represents the initial date and time of contact with a treatment specified in the protocol. Nonetheless, in observational studies, where treatment is not a component, is it acceptable for this variable to be null?	In observational studies where no protocol-specified treatment is administered, it is reasonable to assign a null value to the variable representing the first date/time of exposure to treatment. This accurately reflects the absence of treatment exposure in such studies.
RFXENDTC (Date/Time of Last Study Treatment)	This represents the last date and time of contact with a treatment specified in the protocol. Nonetheless, in observational studies, where treatment is not a component, is it acceptable for this variable to be null?	As it is a variable that is anticipated, removing it entirely would trigger an error from validators like P21. Therefore, our decision was to retain this variable with null values.
ARM (Description of Planned Arm)* **Described below in detail.*	It is representative of the intended arm allocation within the study protocol. However, in our specific study, there was no predefined arm allocation. Consequently, there arises a query regarding the appropriate entry for the ARM variable.	While there is a suggestion circulating across platforms such as CDISC to assign cohort definitions when subjects are distinguished based on their enrollment characteristics, we hold the belief that such an approach may compromise the genuine essence of the variable. Instead, we propose the introduction of a new variable named 'COHORT' within the SUPPDM domain. Concurrently, the ARM variable would designate *'No Investigational Medication*' for the enrolled subjects.

Trial Arm (TA)

Trial arms serve the purpose of enabling comparison between distinct cohorts of participants within a study. These cohorts may receive varying treatments, interventions, exposures, or conditions. In contrast to randomized controlled trials (RCTs), where participants are assigned to different groups by chance, observational studies do not incorporate randomization. Instead, participants are observed, and data is collected based on their pre-existing characteristics or exposures.

As an illustration, in our observational study, subjects were categorized into eight groups:

Asian Healthy	***Non-Asian Healthy***
Asian Mild	***Non-Asian Mild***
Asian Moderate	***Non-Asian Moderate***
Asian Severe	***Non-Asian Severe***

If we adhere to CDISC conformance rules for observational studies in assigning trial arms, the TA domain would be structured as follows:

DOMAIN	ARM	ARMCD	TAETORD	ELEMENT	ETCD	TABRANCH	TATRANS	EPOCH
TA	Asian Healthy	R1	1	Screening	SCR	NA	NA	SCREENING
TA	Asian Healthy	R1	2	Baseline	BAS	NA	NA	BASELINE
TA	Asian Healthy	R1	3	Observational Period	OBS	NA	NA	PERIOD 1
TA	Asian Healthy	R1	4	End of Study	EOS	NA	NA	FOLLOW-UP
TA	Asian Mild	R2	1	Screening	SCR	NA	NA	SCREENING
TA	Asian Mild	R2	2	Baseline	BAS	NA	NA	BASELINE
TA	Asian Mild	R2	3	Observational Period	OBS	NA	NA	PERIOD 1
TA	Asian Mild	R2	4	End of Study	EOS	NA	NA	FOLLOW-UP
TA	Asian Moderate	R3	1	Screening	SCR	NA	NA	SCREENING
TA	Asian Moderate	R3	2	Baseline	BAS	NA	NA	BASELINE
TA	Asian Moderate	R3	3	Observational Period	OBS	NA	NA	PERIOD 1
TA	Asian Moderate	R3	4	End of Study	EOS	NA	NA	FOLLOW-UP

Continuing in this manner would lead to four observations for each group or arm, resulting in a total of 32 observations for the entire TA domain. While this approach is manageable when comparing a minimal number of groups, such as two or three, it becomes unwieldy with a larger sample group. Consequently, we opted for an alternative approach to circumvent this protracted process.

The Trial Arm (TA) domain designed for this study may not align proportionately with the recommendations put forth by CDISC for observational studies. In CDISC guidelines, ARM is typically utilized to delineate specific characteristics of subjects. However, in our approach, ARM was designated as *'No Investigational Medication'*. Instead, we introduced a variable named 'COHORT' within the SUPPDM to comprehensively capture the particulars of the mentioned study groups. These datasets were concurrently developed to encompass both arm and cohort definitions in the study, ultimately proving instrumental in the subsequent stages of analysis.

Trial Arm and Supplementary Demographics Domain of a Cohort Defined Prospective Observational Study:

DOMAIN	ARM	ARMCD	TAETORD	ELEMENT	ETCD	TABRANCH	TATRANS	EPOCH
TA	No Investigational Medication	R1	1	Screening	SCR	NA	NA	SCREENING
TA	No Investigational Medication	R1	2	Baseline	BAS	NA	NA	BASELINE
TA	No Investigational Medication	R1	3	Observation Period	OBS	NA	NA	PERIOD 1
TA	No Investigational Medication	R1	4	End of Study	EOS	NA	NA	FOLLOW-UP

USUBJID	RDOMAIN	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL
1001	DM			COHORT	Cohort	NON ASIAN MODERATE MGD
1002	DM			COHORT	Cohort	NON ASIAN HEALTHY
1003	DM			COHORT	Cohort	ASIAN SEVERE
1004	DM			COHORT	Cohort	ASIAN MILD
1005	DM			COHORT	Cohort	NON ASIAN SEVERE

Supplementary Demographics Domain (SUPPDM)

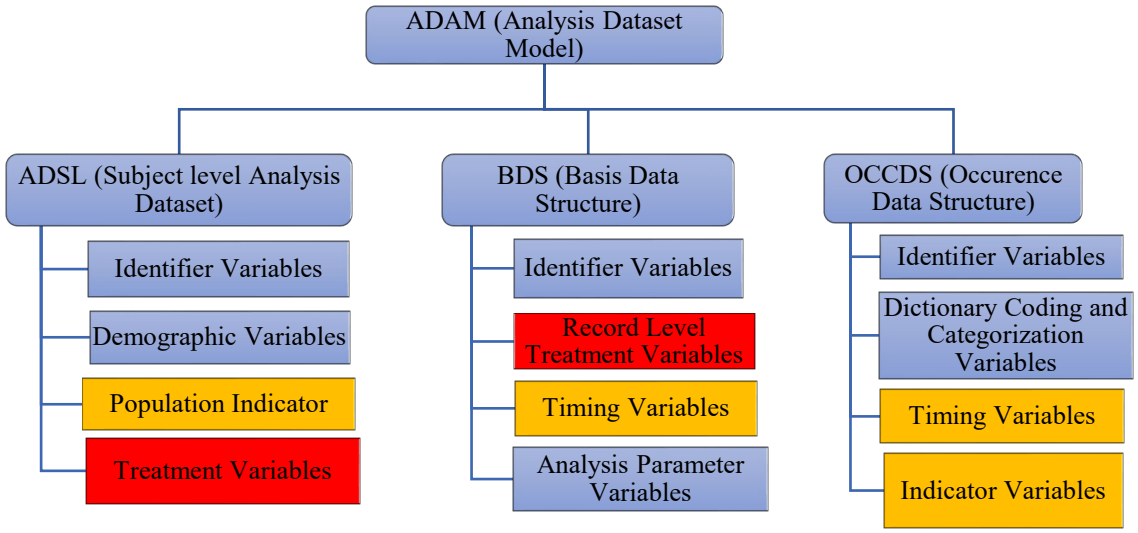
Key Insights

- **Clearly defining Cohorts at an early stage**: Clear and well-defined inclusion and exclusion criteria, along with appropriate stratification factors, are crucial in ensuring that the cohort is appropriately selected and characterized for the objectives of the study. These criteria provide the foundation for meaningful analysis and interpretation of study results.
- **Standardized Data Collection**: Ensure that data collection is standardized across all study sites and timepoints. This includes standardized data collection forms, procedures, and definitions.
- **Harmonization of Variables**: Ensure that variables are defined consistently and adhere to CDISC (Clinical Data Interchange Standards Consortium) standards. This promotes interoperability and facilitates data exchange.

ANALYSIS DATASET MODEL (ADaM)

ADaM stands for Analysis Data Model. It is a set of standards developed by the Clinical Data Interchange Standards Consortium (CDISC) for the creation, analysis, and reporting of clinical trial data. ADaM defines a structured and standardized format for the analysis datasets used in the statistical analysis of clinical trial results.
In the realm of observational studies, we don't adhere to a rigid framework. Thanks to the adaptable nature of ADaMs, we have the flexibility to customize them to suit our specific needs. Nevertheless, there's a considerable amount of ground to cover when it comes to non-interventional studies.

ADaMs are primarily classified into three main data structures: ADSL, BDS, and OCCDS. Analysis datasets corresponding to BDS will adhere to the same set of rules, and the same applies to OCCDS. Therefore, rather than delving into the intricate details of dataset structure, our focus will be on understanding the variables that constitute them. Nevertheless, I still aim to present an overview of the ADaM using the provided diagram. This will assist in evaluating the various variable sections that may not hold significant relevance in the context of observational studies.



Here are the primary variable categories within ADaMs. Keep in mind, this list is not comprehensive. For a more detailed version, consult the ADaM Implementation guide. Variables marked in red are not applicable to observational studies. Those in yellow raise significant concerns as they seem to rely on the red variables for their derivation.

Further, we will delve into some of them in detail.

TIMING VARIABLES

The primary variables included in this category encompass AVISIT, AVISITN, ADTM, ASTDTM, AENDTM and others. These variables are, to some extent, influenced by the VISIT as outlined in the protocol, with particular emphasis on AVISIT and AVISITN.

Visit Handling: The management of visits constitutes a recurring issue confronted by programmers engaged in observational studies. Instances may arise wherein a visit is either absent or necessitates imputation. For instance, observational studies conventionally encompass designated points in time of particular interest, demanding precise vigilance from programmers in orchestrating visit synchronization. It is imperative to underscore that the Statistical Analysis Plan (SAP) serves as an authoritative guide for programmers, delineating the protocols for addressing absent visits or overseeing pertinent time intervals.

Now, let's examine the difficulty we encountered when handling the visit variable in our study. Below is the chronological order of visits as outlined in the assessment schedule specified in our study protocol:

Period	**Observation**						**End Study**
Visit name	Screening	Baseline	Visit 2	Visit 3	Visit 4	Visit 5	Visit 6
Days	-60 to 1	1	90± 10	180± 10	365± 20	1095± 20	1460 ± 20

Subject Visits (SV):

Baseline is missing for Subjects 1001 and 1002.

USUBJID	SVSEQ	VISIT	VISITNUM	VISITDY	EPOCH
1001	1	Screening	1	-1	Screening
1001	2	Day 90	101	90	Period 1
1001	3	Day 180	102	180	Period 1
1002	1	Screening	1	-1	Screening
1002	2	Day 90	101	90	Period 1
1002	3	Day 180	102	180	Period 1
1003	1	Screening	1	-1	Screening
1003	2	Baseline	2	1	Screening
1003	3	Day 90	101	90	Period 1
1003	4	Day 180	102	180	Period 1
1004	1	Screening	1	-1	Screening
1004	2	Baseline	2	1	Screening
1004	3	Day 90	101	90	Period 1
1004	4	Day 180	102	180	Period 1
1005	1	Baseline	2	1	Screening
1005	2	Day 90	101	90	Period 1
1005	3	Day 180	102	180	Period 1
1006	1	Baseline	2	1	Screening
1006	2	Day 90	101	90	Period 1
1006	3	Day 180	102	180	Period 1

Screening is missing for Subjects 1005 and 1006.

Below is the definition of Baseline derivation as outlined in the Statistical Analysis Plan:

> For change from baseline analyses, baseline is defined to be the Day 1 visit for most ocular endpoints. For those endpoints later added with the initial value collected after Day 1, the baseline will be the first measurement or other as stated in each endpoint section.

This clearly illustrates the pivotal role of the Day 1 Visit in establishing the baseline and carrying out corresponding analyses. As per the assessment schedule and the SV (Subject Visits) domain, it was anticipated that participants would attend the site on the designated days. However, if the site neglects to record the Day 1 visit and instead considers the screening visit as the baseline, it gives rise to significant questions in the entire analysis procedure. How would we compute AVISIT and AVISITN? How do we define the baseline? How can we produce the analysis records? The absence of visits, particularly the baseline, in the data introduces various uncertainties and hurdles.

The visits also impact the Population Indicators. In our study, there are two analysis sets that rely on the participants' visits.

- ***Milestone Analysis Set:*** All participants who comply with the entry criteria and have a screening, *baseline* and post-baseline evaluation will be included in the milestone analysis set.
- ***Baseline and Screening Analysis Set:*** This set will consist of the participants who were enrolled in the study and have a *baseline* & screening evaluation.
  It is clear that these analysis subsets rely heavily on the baseline visit, and any disruption to it could potentially jeopardize the entire analysis process.

Proposed Resolution
This is one of the primary data challenges encountered in exploratory studies. Sites may become unresponsive, fail to follow up on time, or may even refuse to rectify the issue from their end. This leaves the analysis team to handle the situation independently. Fortunately, in our case, we received clarification that the issue only occurred with subjects in the first two cohorts, which amounts to nearly 120 subjects. In this scenario, we can consider either the Screening or the Baseline as the reference point, depending on which one is available. However, addressing this issue would necessitate a kind of "first-aid kit" for the study, where variables like ANL01FL (Analysis Flag) have proven to be invaluable in resolving such situations.

TABLE AVISIT MAPPING

USUBJID	PARAM	AVAL	VISIT	VISITNUM	AVISIT	AVISITN	ANL01FL
1001	Vascularity	0	Screening	1	SCR/BAS	3	Y
1001	Vascularity	1	Day 90	101	Day 90	101	Y
1001	Vascularity	2	Day 180	102	Day 180	102	Y
1002	Vascularity	2	Screening	1	SCR/BAS	3	Y
1002	Vascularity	1	Day 90	101	Day 90	101	Y
1002	Vascularity	1	Day 180	102	Day 180	102	Y
1003	Vascularity	0	Screening	1	SCR	1	
1003	Vascularity	0	Baseline	2	SCR/BAS	3	Y
1003	Vascularity	0	Day 90	101	Day 90	101	Y
1003	Vascularity	2	Day 180	102	Day 180	102	Y
1004	Vascularity	1	Screening	1	SCR	1	
1004	Vascularity	2	Baseline	2	SCR/BAS	3	Y
1004	Vascularity	3	Day 90	101	Day 90	101	Y
1004	Vascularity	2	Day 180	102	Day 180	102	Y
1005	Vascularity	1	Baseline	2	SCR/BAS	3	Y
1005	Vascularity	2	Day 90	101	Day 90	101	Y
1005	Vascularity	0	Day 180	102	Day 180	102	Y
1006	Vascularity	3	Baseline	2	SCR/BAS	3	Y
1006	Vascularity	1	Day 90	101	Day 90	101	Y
1006	Vascularity	2	Day 180	102	Day 180	102	Y

The records marked in green were excluded from the analysis.

The designation of AVISIT (Analysis Visit) as 'SCR/BAS' was made to represent both the screening and baseline visits together, as our study's design allowed for this customization. It's important to note that this approach may not be applicable to other studies with different design structures.

POPULATION INDICATOR: The Population Flags were determined based on pre-existing visits, introducing a degree of unreliability to the analysis. For example, the Milestone Analysis Set encompassed all subjects with screening, baseline, and post-baseline evaluations. From this, it was inferred that subjects with both screening and post-baseline evaluations must have also undergone a baseline evaluation. Consequently, a 'Y' flag was assigned to these subjects. While this approach proved effective for our study due to the flexibility of its design, it may not be suitable for more intricate study designs.

TREATMENT VARIABLES AND RECORD LEVEL TREATMENT VARIABLES:

Treatment Variables	TRT01P, TRT01PN, TRT01A, TRT01AN etc.
Record Level Treatment Variables	TRTP, TRTPN, TRTAN, TRTPN etc.

This list of variables is not comprehensive. Please consult the ADaMIG for an extensive inventory of variables.

These particular variables may not pertain to observational studies. Nevertheless, a pertinent question arises: should we retain these variables or exclude them? If we choose to include them, what should be entered? Conversely, if we opt to exclude them, how do we address potential Pinnacle 21 issues, given that some of these variables, such as ARM and TRT01P, are obligatory? In our study, we decided to populate them with the same information as the ARM variables, specifically, *'No Investigational Medication'*.

Key Insights

- **Clear Data Standards and Conventions**: Establish clear and standardized data conventions and standards for data collection, coding, and formatting. This ensures consistency and facilitates smoother data processing**.**
- **Comprehensive Data Mapping**: Ensure thorough mapping of source data to ADaM variables. This includes mapping data from various SDTM domains to the appropriate ADaM domains and variables.
- **Data Cleaning and Validation**: Implement robust data cleaning and validation processes to identify and address discrepancies, outliers, and missing values. This helps maintain data integrity and accuracy.

STATISTICAL ANALYSIS PLAN AND TFL SHELLS

The Statistical Analysis Plan (SAP) and TFL (Tables, Figures, and Listings) Shells are two essential components in the process of conducting and reporting the results of a clinical trial or research study. SAP is usually called as the Bible for Statistical Programmers, which serve as the source for TFL Shells. Any update in former will affect the latter as well. Hence, they are closely related in the following ways:

Guidance for Analysis and Reporting:

- SAP: The SAP guides the statisticians and data analysts on how to perform the statistical analyses, including the specific methods and techniques to be used.
- TFL Shells: TFL Shells provide a structured format for organizing and formatting the tables, figures, and listings that will be included in the final study report.

Interdependency:

- SAP: It serves as a reference point for validating that the analyses conducted are in line with the pre-defined plan.
- TFL Shells: They are used to review and validate the actual results generated to ensure they align with the predefined structure.

In observational studies, due to their exploratory nature, endpoints are often unclear, leading to less refined TFL shells. Additionally, the SAP lacks documentation on handling missing data, which poses challenges in the analysis process. The ultimate aim for a programmer is to generate TFLs for the Clinical Study Report (CSR). Unclear TFLs necessitate frequent redevelopment, leading to wastage of time and resources. In our study, several parameters were overlooked in the SAP, consequently affecting the corresponding TFL Shells.

An instance of this is Tear Break Up time, which is a crucial parameter in our research. The SAP did not offer any direction on how to calculate it, and the TFL Shells did not cover this aspect either. However, during the analysis, the team concluded that it was necessary to take an average of all the readings rather than considering individual recordings. Although such modifications are typical, even in intervention-based studies, this specific case highlights one of several parameters/tests that necessitated adjustments at a later stage, leading to redundant efforts for the programmers.

Key Insights

- **Undocumented Parameters:** As in our case, important parameters like Tear Break-Up time were not accounted for in the SAP or TFL Shells. This highlights the need for a more comprehensive and detailed SAP, especially in observational studies where endpoints might not be as clear-cut.
- **Derivations and Interpretations:** It's not uncommon for certain parameters or tests to require derivations or adjustments during the analysis phase. However, it's crucial that these adaptations are clearly documented and communicated to avoid confusion and duplication of work.
- **Documentation and Reproducibility:** Clear documentation of data preparation, analysis steps, and code is vital for reproducibility. Inadequate documentation can hinder the ability of others to replicate the study.

In summary, the SAP establishes the plan for how data will be analyzed, while TFL Shells provide the structured format for presenting the actual results of the analyses. The SAP informs what needs to be done, and the TFL Shells are used to document and present the outcomes of those analyses. Both documents work together to ensure transparency, consistency, and rigor in the reporting of study results. SAP needs to be more robust and detailed to account for the parameters requiring the adjustments. It serves as a guiding light for the statisticians and statistical programmers, any hindrance with it will mess up the whole study.

CONCLUSION

Cohort-defined observational studies play a pivotal role in epidemiological research by offering a clear temporal sequence between exposures and outcomes, thereby illuminating causal relationships. They excel in representing real-world scenarios and diverse populations, underscoring their importance in advancing medical knowledge and guiding public health interventions. Despite the challenges they present, particularly for statistical programmers, these studies bring substantial benefits. Collaborative efforts, such as those seen in communities like PHUSE Working Groups, are making significant contributions, and clearer guidelines are anticipated in the future. This will help individuals anticipate challenges and discover effective workarounds. Effective analysis hinges on the collaboration between data managers, statisticians, and statistical programmers, alongside specialized training, and official guidance. This field offers a fertile ground for programmers to explore standards. Rather than seeing challenges as roadblocks, they should be viewed as opportunities to find solutions.

REFERENCE

1. FDA, “Study Technical Conformance Guide” (v5.4, June 2023) https://www.fda.gov/media/153632/download
2. Considerations for Using CDISC Standards in Observational Studies | CDISC
3. Considerations and Examples of Using SDTM for Observational Studies - sections - Considerations and Examples for using SDTM for Observational Studies - Wiki (cdisc.org)
4. PP24.pdf (lexjansen.com)
5. SDTMIG | CDISC
6. ADaM | CDISC

ABBREVIATION

AVISIT	Analysis Visit
AVISITN	Analysis Visit Numeric
ASTDTM	Analysis Start Datetime
AENDTM	Analysis End Datetime
ADTM	Analysis Datetime
ANL01FL	Analysis Flag 01

CONTACT INFORMATION

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

Author Name: Vikas Shokeen
Company: Novartis
Address: NKC, Hyderabad, India
Email: vikas.shokeen@novartis.com