Rapport Intermédiaire – Tableau 1 Baseline

Nom et Prenom : Abdo MIRGAN

M1 Intelligence des Données en Santé

1. Introduction

Ce rapport présente les étapes complètes de traitement, de nettoyage et de construction du Tableau 1 dans le cadre du Data Challenge portant sur un essai clinique. L'objectif est de décrire précisément la population à l'inclusion (baseline), conformément aux standards (SDTM).

2. Jeux de données SDTM utilisés

Le fichier d'annotation SDTM a permis d'identifier l'origine des variables et les formulaires cliniques utilisés pour alimenter les différentes bases. Les données du Tableau 1 provisoire proviennent principalement :

- du domaine *DM* (démographie) issue du formulaire **CTN Demographics Form**
- du domaine *SC* (caractéristiques issues du formulaire **Subject Characteristics**)
- du domaine *VS* (signes vitaux enregistrés dans les formulaires **Vital Signs** et **Physical Examination**)

Ce document d'annotation a servi de guide pour comprendre comment les informations collectées dans les formulaires CRF ont été standardisées dans les domaines SDTM et pour sélectionner les variables pertinentes.

3. Pipeline méthodologique

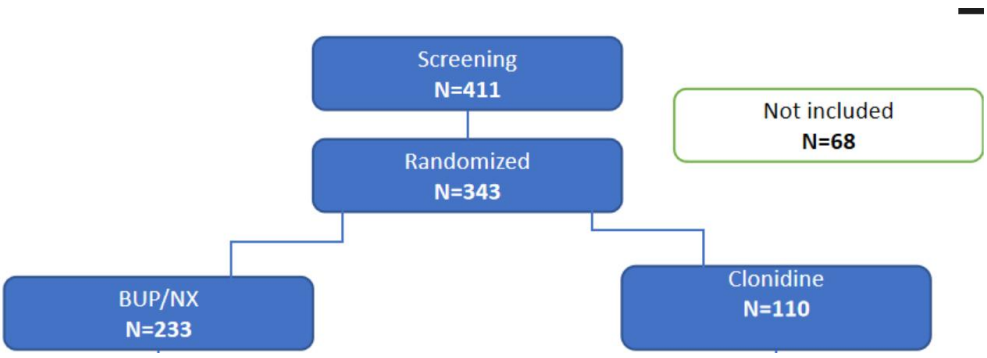
Le traitement suit les étapes suivantes :

- L'analyse repose sur l'utilisation de Polars pour le traitement principal des données, pandas pour un besoin ponctuelles et TableOne pour générer le Tableau 1.
- Nous avons importé et concaténé les données démographiques, les caractéristiques associées et les signes vitaux issus des dossiers CTN0001 et CTN0002 afin d'obtenir des jeux de données completes pour effectuer des traitements.

3.1 Etape de traitement des variables démographique

- Dans la tables dm.csv , nous avons conservé uniquement les variable démographiques nécessaires au Tableau 1 (USUBJID, ARMCD, AGE, SEX et RACE). Cette première table servira de base à laquelle seront ajoutées les autres caractéristiques via des jointures avec les autres datasets. Que l'on appelleras dataframes finales.

- La cohérence de ce dernier a été vérifiée en confirmant la présence de 411 sujets à l'inclusion, conformément au flowchart initial.



3.2 Etape de traitement des variables démographique supplementaire.

- Pour compléter les caractéristiques démographiques, nous avons extrait depuis la table *sc.csv* (domaine SC) les variables décrivent le profil socio-économique des participants. .

- Puis avions joint la table contenant seulement ces variables démographique supplémentaires à notre dataframe finsales via leur indentifiant unqiue USUBJID attribué dans le contexte de l'essaie clinique.

- Au final nous avons un dataframe avec les variables démographiques suivante : identifiant unique , bras de traitements , age , sexe , race , nombre d année d'etude , le status marital , le status professionnels .

- L age et le nombre d année d'etude ont été convertis en float

Autre étape effectuer : **Recuperer le poids et la taille dans la table vs.csv pour creer une colonne BMI et l'ajouter a notre dataframe finale.**

Pour obtenir l'indice de masse corporelle (BMI), nous avons d'abord isolé **HEIGHT** et **WEIGHT** depuis les données des signes vitaux (VS). Comme ces mesures sont

exprimées en **pouces** (inches) et **livres** (pounds), elles ont été converties en unités SI , puis ont à calculé le BMI

Une fois le BMI obtenu, nous avons supprimé HEIGHT et WEIGHT pour ne conserver que le **BMI**, qui à ensuite été jointes au tableau final.

3.3 Etape de traitement des variables caractéristiques des signes vitaux .

Pour les signes vitaux, nous avons extrait et harmonisé les mesures issues du domaine **VS**, en procédant par étapes structurées :

- **Sélection des tests pertinents**
  Conservation uniquement des variables standard de signes vitaux :
  DBP, SBP, PULSE, RESP, TEMP, ainsi que **HEIGHT** et **WEIGHT** (utiles pour le BMI).
- **Restriction au baseline**
  Filtre sur **VISITNUM = 0** afin de ne garder que les valeurs d'inclusion.
- **Contrôle des formulaires**
  Utilisation exclusive des mesures provenant des formulaires cliniques appropriés : **VITAL SIGNS FORM** et **PHYSICAL EXAMINATION FORM**, identifiés via **VSCAT**.
- **Standardisation de la position**
  Filtre sur **VSPOS = "SITTING"** pour éviter la duplication des mesures prises en positions différentes (SITTING/STANDING).
- **Gestion des doublons**
  Plusieurs patients avaient plusieurs mesures pour le même test à baseline.
  → Solution : agrégation par **moyenne** (group_by USUBJID + VSTESTCD).
- **Pivotage long → wide**
  Transformation du tableau pour obtenir **une seule ligne par patient**, avec une colonne par signe vital.
- **Conversion d'unités**
  Conversion de la température en **°C** depuis les degrés Fahrenheit :
- **Nettoyage final**
  Sélection et renommage des colonnes pour constituer le bloc final de variables physiologiques :
  - Pression diastolique
  - Pression systolique
  - Pulsations/min
  - Cycles respiratoires/min
  - Température en °C
- Ajout des variables physiologiques à la base démographique-socio-économique (data frame finale) , par jointure.

4. Production du Tableau 1 (TableOne)

4.1 Construction du Tableau 1

Pour générer le Tableau 1 comparant les caractéristiques des participants selon le bras de traitement (**ARMCD**), nous avons procédé comme suit :

- **Conversion en format compatible**
  Le package *TableOne* ne supportant pas Polars, le dataframe finale a été convertie en DataFrame **pandas**

- **Définition des variables**

  - **Catégorielles** : SEX, RACE, MARITAL, EMPLOY30
  - **Continues** : AGE, EDUCYSR, BMI, pressions artérielles, fréquence cardiaque, fréquence respiratoire, température

- L'objet **TableOne** a ensuite été créé en spécifiant les variables à afficher (categorielles + Continues) et la variable de stratification (Bras de traitement)
- **Sortie finale**
  Le tableau obtenu synthétise :
  - la distribution des variables démographiques, socio-économiques et physiologiques,
  - **stratifiée par bras de traitement**,
  - avec effectifs, proportions, moyennes et écarts-types.

Tableau 1 partielle : en bas de page

4.2 Tableau 1 partielle

			Grouped by ARMCD			
		Missing	Overall	BUPNAL	CLON	SCRFAIL
n			411	233	110	68
SEX, n (%)	F		128 (31.1)	72 (30.9)	38 (34.5)	18 (26.5)
	M		281 (68.4)	161 (69.1)	72 (65.5)	48 (70.6)
	U		2 (0.5)	0 (0.0)	0 (0.0)	2 (2.9)
RACE, n (%)	BLACK, AFRICAN AMERICAN, OR NEGRO		120 (29.2)	71 (30.5)	35 (31.8)	14 (20.6)
	None		2 (0.5)	0 (0.0)	0 (0.0)	2 (2.9)
	OTHER		24 (5.8)	12 (5.2)	6 (5.5)	6 (8.8)
	SPANISH, HISPANIC, OR LATINO		87 (21.2)	45 (19.3)	19 (17.3)	23 (33.8)
	WHITE		178 (43.3)	105 (45.1)	50 (45.5)	23 (33.8)
MARITAL, n (%)	DIVORCED		63 (15.3)	38 (16.3)	17 (15.5)	8 (11.8)
	LEGALLY MARRIED		74 (18.0)	47 (20.2)	15 (13.6)	12 (17.6)
	LIVING WITH PARTNER/COHABITATING		38 (9.2)	20 (8.6)	12 (10.9)	6 (8.8)
	NEVER MARRIED		188 (45.7)	109 (46.8)	51 (46.4)	28 (41.2)
	None		6 (1.5)	0 (0.0)	0 (0.0)	6 (8.8)
	SEPARATED		32 (7.8)	14 (6.0)	12 (10.9)	6 (8.8)
	WIDOWED		10 (2.4)	5 (2.1)	3 (2.7)	2 (2.9)
EMPLOY30, n (%)	FULL TIME (35+ HRS/WK)		133 (32.4)	73 (31.3)	44 (40.0)	16 (23.5)
	HOMEMAKER		14 (3.4)	8 (3.4)	4 (3.6)	2 (2.9)
	IN CONTROLLED ENVIRONMENT		1 (0.2)	1 (0.4)	0 (0.0)	0 (0.0)
	None		7 (1.7)	0 (0.0)	0 (0.0)	7 (10.3)
	PART TIME (IRREGULAR DAYWORK)		45 (10.9)	28 (12.0)	9 (8.2)	8 (11.8)
	PART TIME (REGULAR HOURS)		17 (4.1)	11 (4.7)	4 (3.6)	2 (2.9)
	RETIRED/DISABILITY		11 (2.7)	3 (1.3)	6 (5.5)	2 (2.9)
	STUDENT		8 (1.9)	4 (1.7)	2 (1.8)	2 (2.9)
	UNEMPLOYED		175 (42.6)	105 (45.1)	41 (37.3)	29 (42.6)
AGE, mean (SD)		68	38.0 (10.1)	37.4 (10.5)	39.2 (9.3)	nan (nan)
EDUCYRS, mean (SD)		6	12.5 (2.1)	12.6 (2.0)	12.7 (2.3)	11.8 (1.9)
BMI, mean (SD)		34	25.1 (4.9)	24.8 (5.0)	25.6 (4.6)	25.7 (5.3)
Pression_diastolique_mmHg, mean (SD)		28	79.9 (34.0)	80.9 (43.0)	80.0 (10.8)	74.9 (11.9)
Pression_systolique_mmHg, mean (SD)		28	121.1 (16.5)	120.5 (15.9)	124.4 (17.1)	116.9 (16.8)
battement_minute, mean (SD)		29	75.7 (12.1)	76.3 (11.8)	75.2 (12.7)	74.0 (12.0)