



Database Tools Needed for Personalized Study Participant Risk Surveillance

PHUSE Connect 2023 – Orlando, Florida



The difference is in the data.™

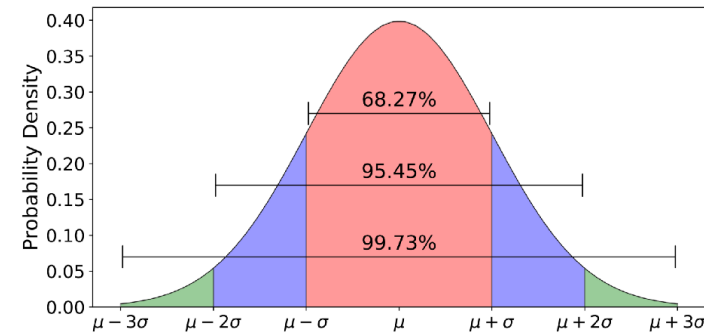
Current Best Practice for Clinical Trials

- Participant screening prior to enrolment
- Safety data from pre-clinical trials
- Good Clinical Practice (GCP) adherence
- Monitoring of physical and behavioural traits in participants
- Signs and signals from lab work data 

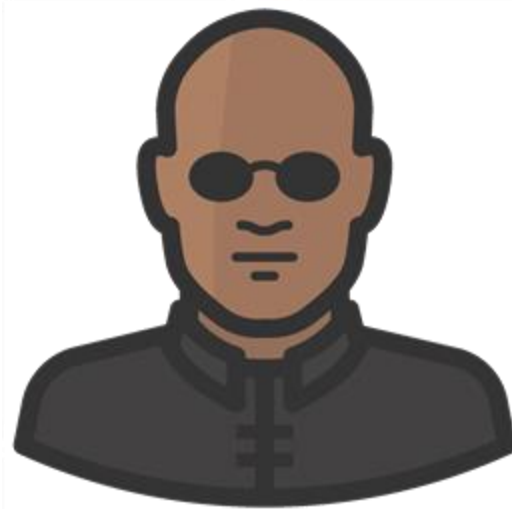
Ex. PCO² elevation compared to a non-trial normalized population is an indicator of a potentially pending adverse event

Cautionary and Warning Signals From Labs

- GCP has triggers for communicating atypical values to investigators and site staff
- Known atypical lab values are considered in decisions to:
 - Keep the participant in the study
 - Remove the participant from the study
 - Advise the participant to seek medical care
 - Provide clinical/medical care to the participant
- Normal, borderline and atypical values are derived from data from thousands and thousands of people
- All these data points populate a normal distribution for the lab values



What if I Told You



**Normal Distributions Neglect Many
Genotypic and Phenotypic Groups?**

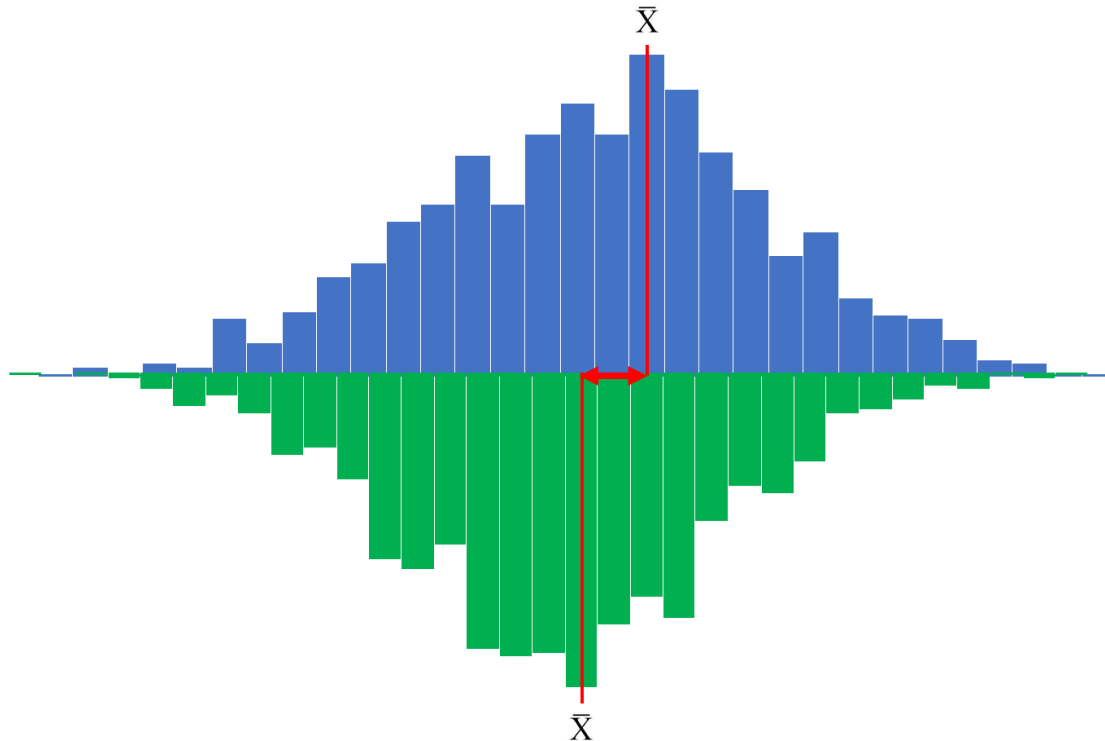
Literature on Within Groups Distributions Outside Full Population Distributions

- Kuś, Aleksander et al. "Variation in Normal Range Thyroid Function Affects Serum Cholesterol Levels, Blood Pressure, and Type 2 Diabetes Risk: A Mendelian Randomization Study." *Thyroid : official journal of the American Thyroid Association* vol. 31,5 (2021)
- Chang, Sheng-Nan et al. "Sex, racial differences and healthy aging in normative reference ranges on diastolic function in Ethnic Asians: 2016 ASE guideline revisited." *Journal of the Formosan Medical Association*
- Varma, Adarsh et al. "African Americans Demonstrate Significantly Lower Serum Alanine Aminotransferase Compared to Non-African Americans." *Journal of racial and ethnic health disparities* vol. 8,6 (2021): 1533-1538.
- Cogswell, Mary E et al. "Estimated 24-Hour Urinary Sodium and Potassium Excretion in US Adults." *JAMA* vol. 319,12 (2018): 1209-1220.

A General Summary of the Literature

- Phenotypic, geographic, SES, medical history, current medications and supplements influence many baseline physiological processes
- Participant provided personal information can be used in identifying groupable traits and characteristics that deviate from broader population normal curves
- Study participants with certain identifiable traits can have lab values distributed much different than the broader population
- Specific harm events can be signaled with less latency and greater validity when normalized distributions consider within groups effects

Normally Distributed $\neq$ The Same Distribution



*Data fabricated to create an abstract visual representation

CDISC Standards for Clinical Trial Data Collection

Data uniformity by design



**Study Data Tabulation Model
Implementation Guide: Human Clinical Trials**
Version 3.4 (Final)

Developed by the
CDISC Submission Data Standards Team

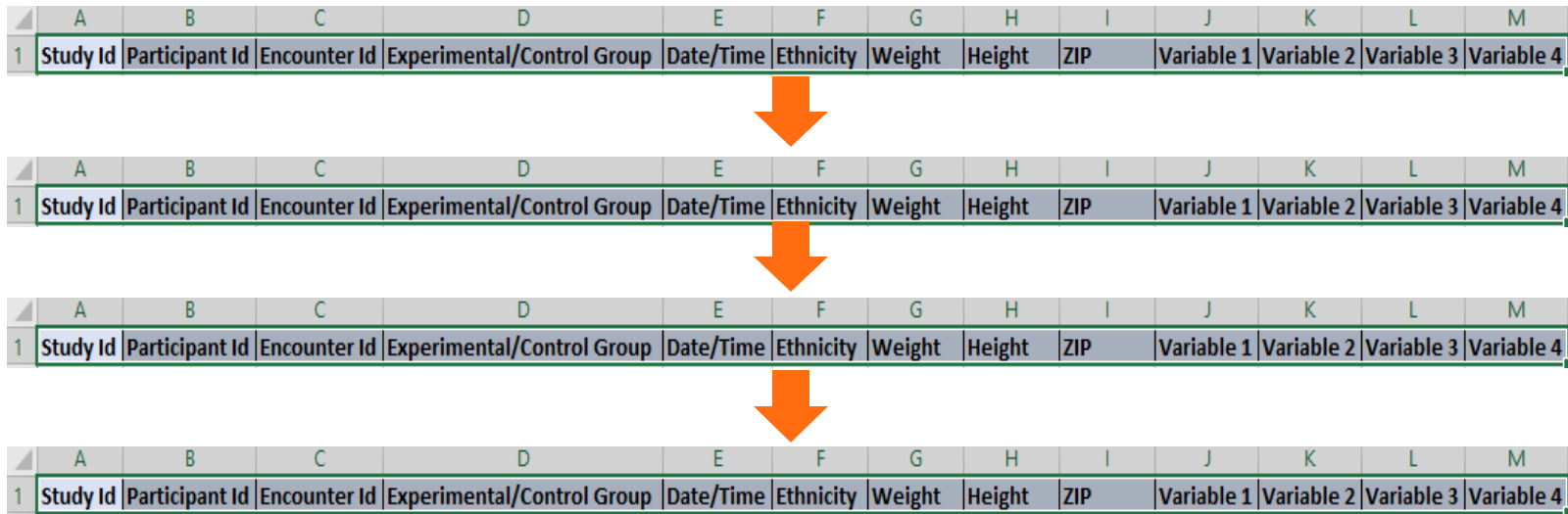


**Clinical Data Acquisition Standards
Harmonization Model**
Version 1.2 (Final)

Prepared by the
CDISC CDASH Team

In-House Data Repositories & Stacked Data

Uniform datasets allow for the data from many related and unrelated trials to be placed one atop of another in a massive but integrated file



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Study Id	Participant Id	Encounter Id	Experimental/Control Group	Date/Time	Ethnicity	Weight	Height	ZIP	Variable 1	Variable 2	Variable 3	Variable 4

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Study Id	Participant Id	Encounter Id	Experimental/Control Group	Date/Time	Ethnicity	Weight	Height	ZIP	Variable 1	Variable 2	Variable 3	Variable 4

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Study Id	Participant Id	Encounter Id	Experimental/Control Group	Date/Time	Ethnicity	Weight	Height	ZIP	Variable 1	Variable 2	Variable 3	Variable 4

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Study Id	Participant Id	Encounter Id	Experimental/Control Group	Date/Time	Ethnicity	Weight	Height	ZIP	Variable 1	Variable 2	Variable 3	Variable 4

* As variance in collected fields and added fields may exist, some manipulation of the data may be necessary

I've Got a Huge Dataset-What Do I do With It?

- Think of the **Normal** in normal distribution
 - Limit the sample to control group participants
 - The effect of the study medication should not be included
- Group the control participants by ALL characteristics in the dataset
 - This should include all demographic, geographic, health history, current medications and others

```

SELECT
    COUNT([Person_Id])
  , COUNT([Encounter_Id])
  , [Sex]
  , [Ethnicity]
  , [Age] → Use a case statement to group to age bands
  , [Country_of_Residence]
  , [Height] → Use a case statement to group to feet or meters
  , [BMI]
  , [Smoking_History]
  , Case statement on presence or absence of disease/disorder/trauma history groups*
  , Case statement on presence of absence of classes of medications/supplements taken**

FROM
    [Datawarehouse].[dbo].[Big_Honking_Table]

WHERE
    [Control/Experimental] = 'Control'

GROUP BY
    [Sex]
  , [Ethnicity]
  , [Age] → Use a case statement to group to age bands
  , [Country_of_Residence]
  , [Height] → Use a case statement to group to feet or meters
  , [BMI]
  , [Smoking_History]
  , Case statement on presence or absence of disease/disorder/trauma history groups*
  , Case statement on presence of absence of classes of medications/supplements taken**
  
```

Logical Groupings and Case Statements

- Too many bins will yield groups too small for sufficient power
- Grouping on traits that are logically similar may provide sufficient power across groups
- Consult the literature on groupings widely used groupings
ex. Classes of drugs the participant is taking, ICD10 groupings...

```

CASE
WHEN [Age] <1 THEN 'Infant'
WHEN [Age] IN (1,2) THEN 'Toddler'
WHEN [Age] IN (3,4) THEN 'Preschool'
WHEN [Age] BETWEEN 6 AND 12 THEN 'School Age'
WHEN [Age] BETWEEN 13 AND 17 THEN 'Adolescent'
WHEN [Age] BETWEEN 18 AND 21 THEN 'Young Adult'
WHEN [Age] BETWEEN 22 AND 64 THEN 'Adult'
WHEN [Age] >= 65 THEN 'Senior'
ELSE 'Unknown'
END AS [Age_Band]

```

Important Considerations for Groupings

- Some combinations will not reach 31 cases based on logic for the broader sample
- Use a SORT BY ASC on the case counts to quickly identify small combinations of traits that make a group
 - Use of a HAVING clause (ex. `HAVING COUNT(Person_Id) <=30`) may help
- Logical groupings on small n combinations should be considered for a second pass of the data
 - If no logical groupings can be found defaulting to the values across the entire sample may be appropriate

WAIT!!!

Didn't you say this was about normal ranges and defining atypical values in the population of a study?!?!?!?

This is all database stuff.....

Parametric or Non-Parametric



That is the Question....

Defining of Measures to be Applied

- It's critical to consider the shape and spread of your data
- In the first look at how the variables aggregate simple functions in SQL work well
 - Histograms for all variables would be overwhelming
 - AVG(), MEDIAN(), MIN(), MAX(), KURTOSIS()
- Personally, I sort by observed kurtosis (but make it absolute for ease of viewing)
 - Many consider a kurtosis of -1 to +1 to indicate a normal distribution
- Others use Pearson's Coefficient of Skewness: $3 * (\text{average} - \text{median}) / \text{stddev}$
- Using Pearson's Coefficient of Skewness 0 = parametric, further from 0 = more skewed

Creation of the Calculated Values Table

- The results on testing for parametricity should be included for reference
- For each group the mean, median, standard deviation, 2.3rd, 15.5th, 84.1st and 97.7th percentiles.
- Each case should now be calculated, parametric between standard deviations, nonparametric between the ranges listed above.
- Cases can now be assigned as normal -1 SD to +1 SD, borderline +/-2 SD to +/-1 SD and atypical <-2 SD or > 2 SD (or with the nonparametric proxies if the data are not normally distributed)

Putting it All Together

- You put the lime in the coconut
 - That's a different talk

Putting it All Together

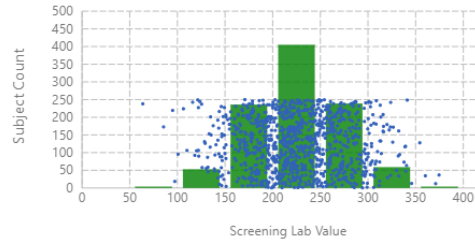
- You now have a derived table from the large dataset with rows for each participant
- The calculated table has all the distinct groupings, calculated values and a range based on the values specific within groups
- At this point any new data arriving can be joined with this table upon import
- Using SSIS or other similar tools, specific cases with lab values in the borderline and atypical ranges can trigger email to site staff and or the PI

Application of These Data

- A tool for greater specificity in detecting potential warnings with greater attention to characteristics of participants
- Tables that can be automatically joined on to incoming data provides rapid feedback and cautions to site staff and the PI
- Data used to create the calculations table should be in the repositories of companies/institutions that are analyzing CDISC data
- No expensive software is required beyond general database tools such as SQL

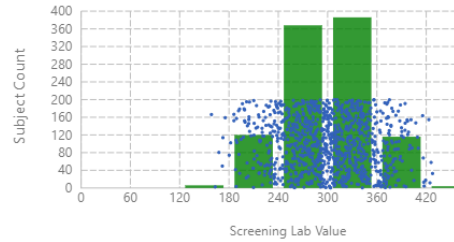
As a visual

Global Population All Blood Types



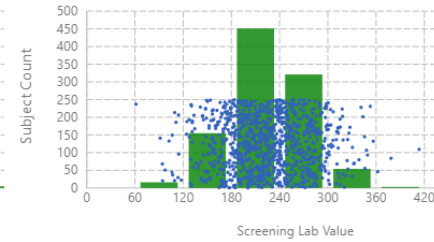
Values less than 131 are statistically defined as outliers within the normal curve and may indicate further investigation or trigger an alert.

Blood Type A



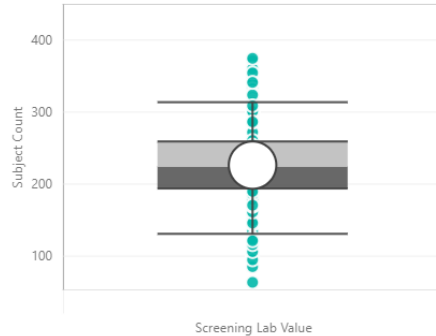
Values less than 200 are statistically defined as outliers within the normal curve and may indicate further investigation or trigger an alert.

Blood Type B

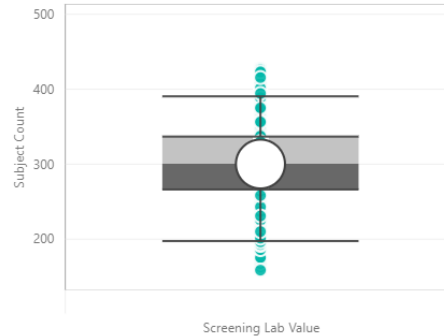


Values less than 128 are statistically defined as outliers within the normal curve and may indicate further investigation or trigger an alert.

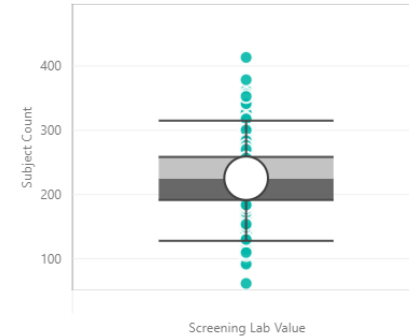
Global Population All Blood Types



Blood Type A



Blood Type B





Thank you!
Any questions?



The difference is in the data.™