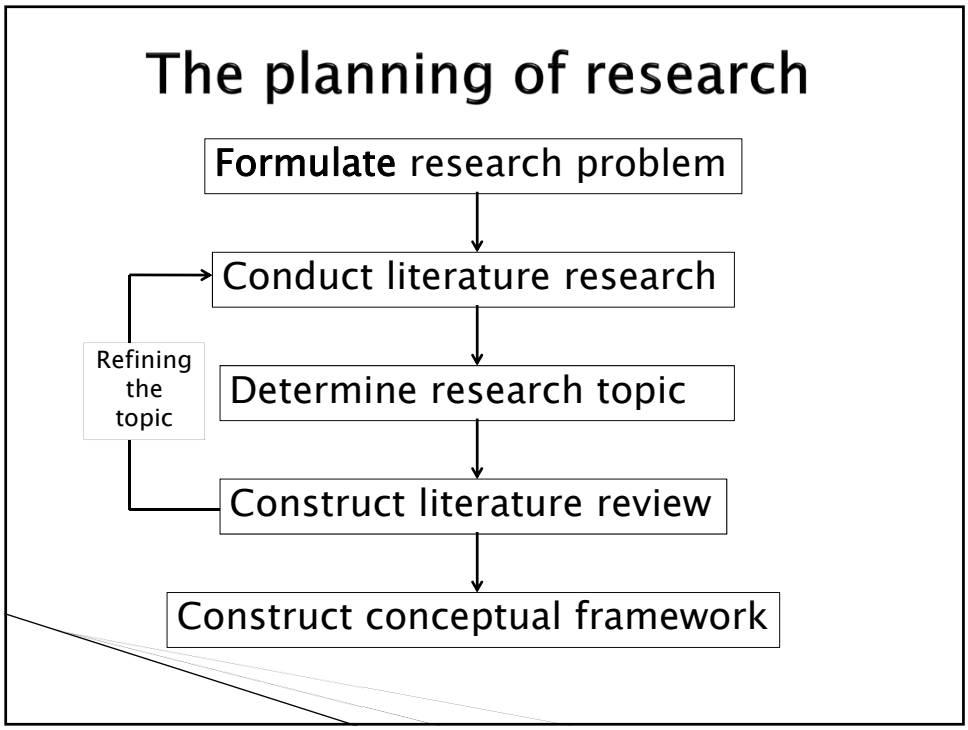
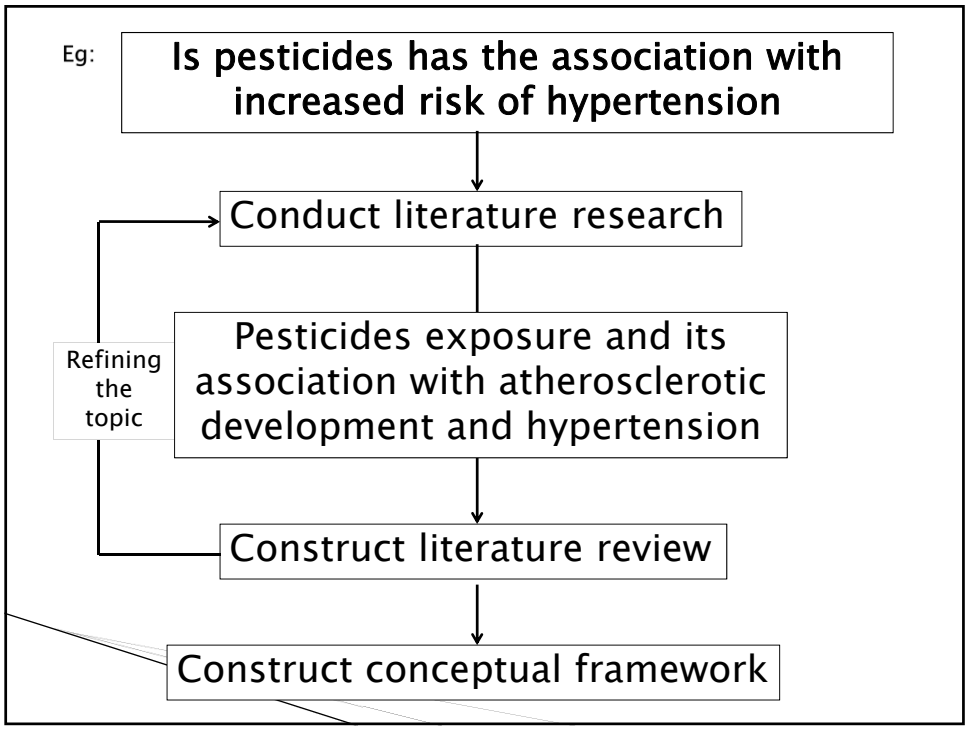
PLANNING FOR DATA COLLECTION & STATISTICAL ANALYSES

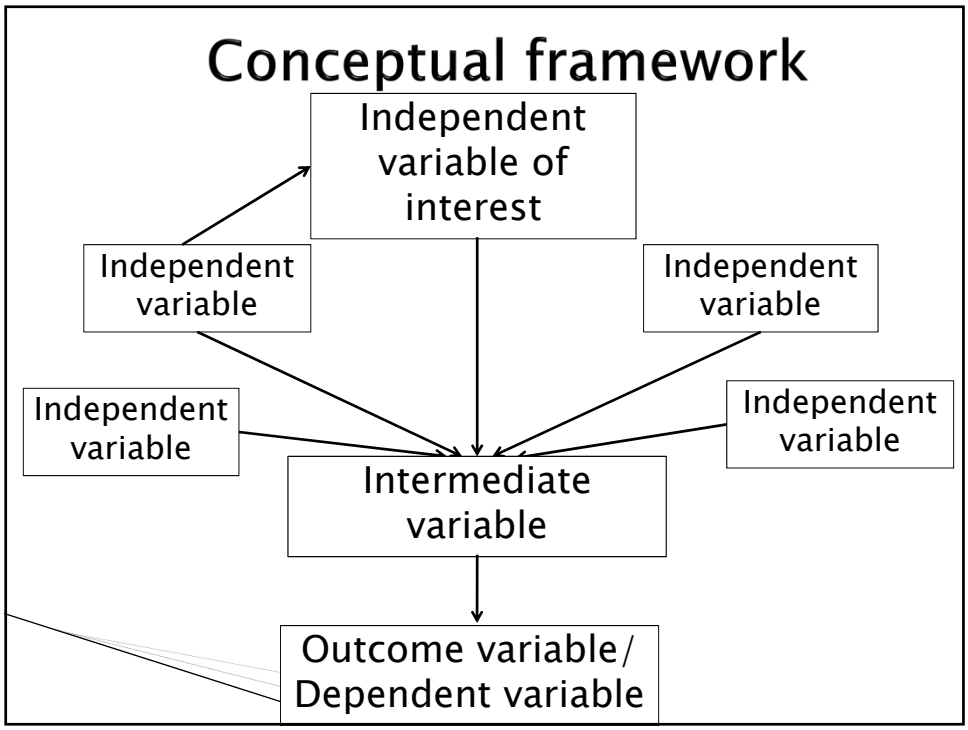
Dr. Niza Samsuddin
Department of Community Medicine
Kulliyyah of Medicine
International Islamic University Malaysia

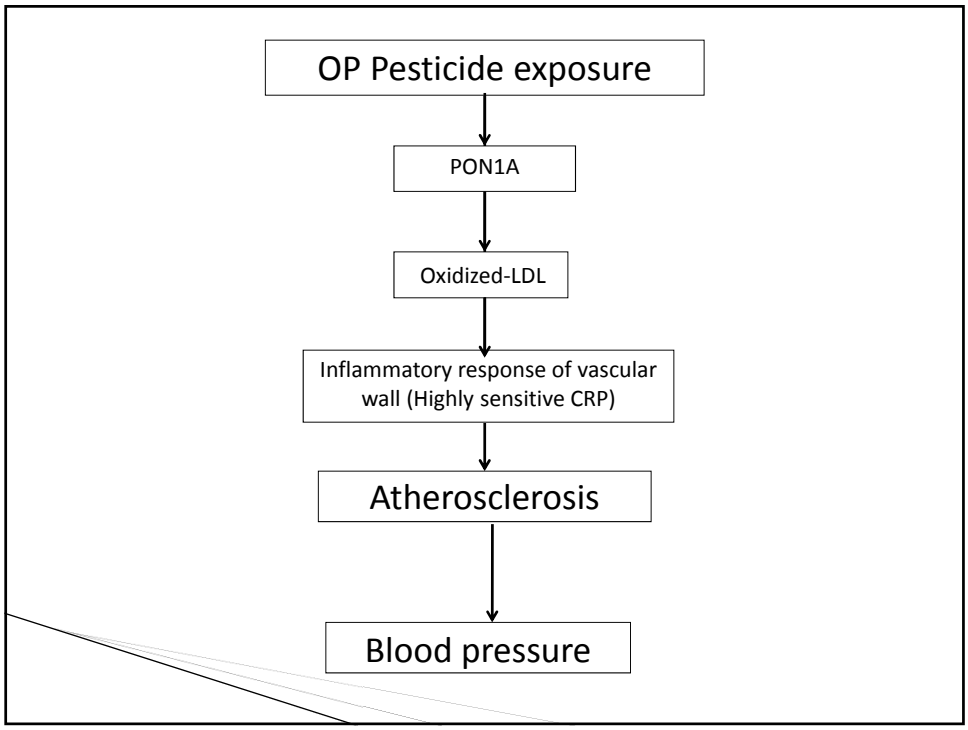
Learning outcome

- At the end of the session, students should be able to:
  - Describe the variables used in the research
  - Choose the best method to collect data
  - Validate the research instruments





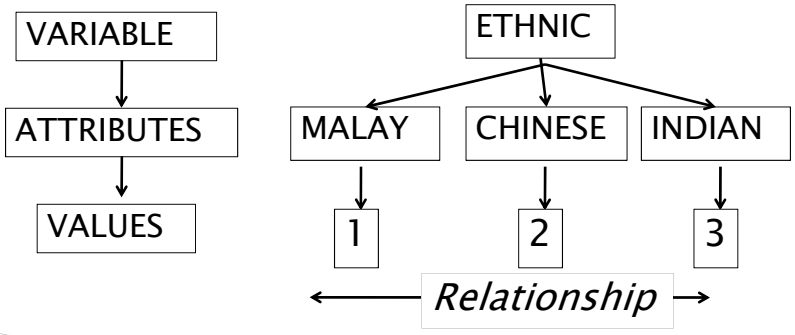




DESCRIPTION OF VARIABLES

Levels of measurement

- The level of measurement refers to the relationship among the values that are assigned to the attributes for a variable



Levels of measurement/ scales of measure

- The types are nominal, ordinal, interval, and ratio.
  - Nominal
  - Ordinal
  - Interval
  - Ratio

- **Nominal measurement**
  - The numerical values just name the attribute accordingly.
  - Weakest and most basic level of measurement
  - Also called categorized or frequency data
  - The attributes must mutually exclusive

- Ethnic
  1=Malay,
  2=Chinese,
  3= India

- Occupation
  1= government servant,
  2= working in private sector,
  3=housewife,
  4=student

- **Ordinal measurement**
  - the attributes can be rank-ordered.
  - The interval between values is not interpretable in an ordinal measure.
  - Can make statement about the relative magnitude (size or scores)

Likert -scale

- Education level
  1= No formal education
  2= primary level
  3= secondary level
  4= tertiary level
- Stage of cancer
  1= stage I
  2= stage II
  3= stage III
  4= stage IV

- **interval** measurement
  - Measurement are on a standardized scale of fixed units separated by **equal** distance.
  - compute an average of an interval variable, temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80. But note that in interval measurement ratios don't make any sense - 80 degrees is not twice as hot as 40 degrees (although the attribute value is twice as large).
  - Eg: standardized personality, IQ scale

- **ratio** measurement
  - strongest level of measurement
  - Measurements starts at an absolute zero point.
  - Represent physical quantity
  - Eg: time, weight, pressure, length, ages

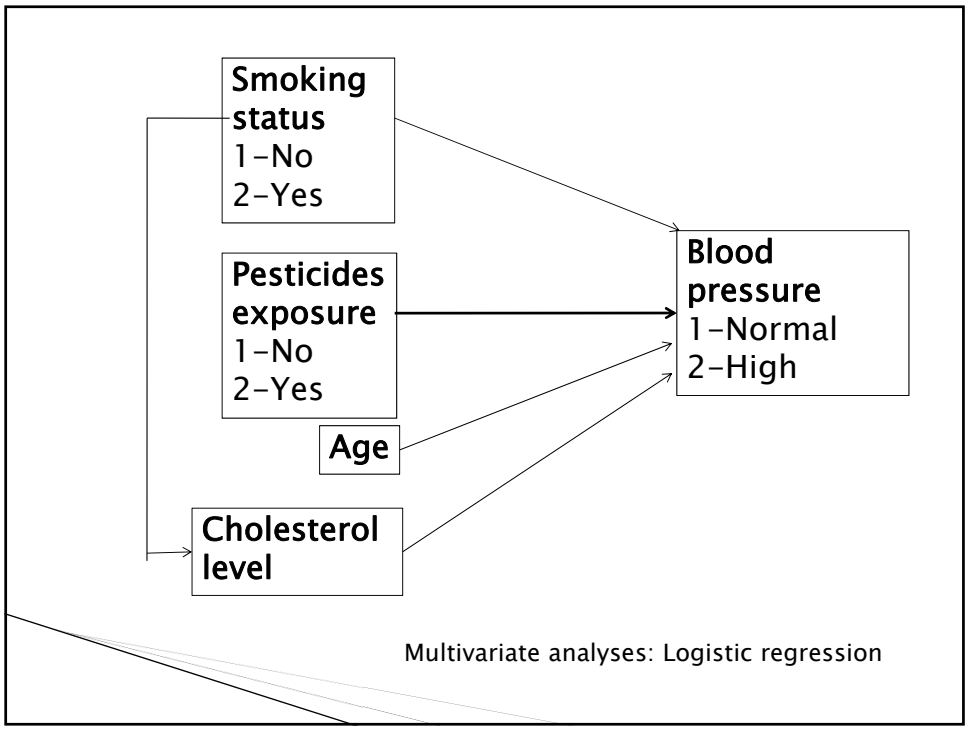
- Why is levels of measurement is important?
  - Interpretation of data
  - Provide short codes for a longer variable name
    - Eg: Family history of premature death due to CVS disease
  - Planning in data analyses
    - For a variable with a nominal measure, the suitable description of the data is in the form of percentage /proportion, not in the form of measures of central tendency (mean, median)
  - Running the statistical analyses
    - Multivariate analyses can only be performed for variables with values for each attributes

Descriptive statistics and graphs for the difference levels of measurements

Levels of measurement	Data properties	Descriptive statistic	Graphs
Nominal	Discrete	Frequencies/ Percentage	Bar/Pie
Ordinal	Discrete Ordered/ rank	Frequencies/ Percentage Min/Max/Range Median/Percentiles	Bar/Pie
Interval	Continuous Metric Equal distance	Min/Max/Range Median/Percentiles Mean/SD	Bar/Pie Stem and Leaf Histogram
Ratio	Continuous Metric Meaningful 0	Min/Max/Range Median/Percentiles Mean/SD	Stem and Leaf Histogram Box & Plot

Bivariate analyses for difference levels of measurements

Variable 1	Variable 2	Statistical test
Nominal	Nominal/Ordinal	Chi-square test
Ordinal	Nominal/Ordinal	Chi-square test
Interval	Nominal/Ordinal	Independent t-test Paired t-test Mann-Whitney U test
Ratio	Nominal/Ordinal	Independent t-test Paired t-test Mann-Whitney U test
Ratio	Ratio/Interval	Person Correlation test Spearmen Correlation test



Multivariate analyses: Logistic regression

Construct dummy table based on the research objectives.

One of the objectives:

- To measure the association between the demographic characteristics and blood pressure among workers exposed to pesticides

	Normal Blood Pressure	High blood pressure	P value
Age	Mean(SD)	Mean (SD)	Independent t-test
Gender Male Female	 N(%) N(%)	 N(%) N(%)	Chi -square test
Income	Mean(SD)	Mean (SD)	Independent t-test
Pesticides exposure	N(%)	N(%)	Chi -square test

- Literature review revealed that after certain age the risk of getting CVS disease is higher than the normal population. What is your approach in the data collection?

- Age:
  - ❑ >55 years
  - ❑ <55 years

- Age:
  _____ years

- Always choose the highest hierarchy levels of measurement during data collection
  - higher levels of measurement (interval or ratio)
  - lower levels of measurement (nominal or ordinal)

- Can change to ratio measurements when required and justified

Variables definitions (Operational definitions)

List out the different types of variables

Dependent Variable (outcome, response, criterion, or explained variable)

- The variable being measured as an outcome

Independent Variable (explanatory, predictor, or manipulated variable)

- A variable that is selected or controlled by the researcher, to determine its relationship to the observed outcome of the research

Variables definitions (Operational definitions)

Intervening (Mediating) / Intermediate/ process Variables:

- A concept that attempts to explain the relationship between the independent and dependent variables (Baron & Kenny, 1986).
- Explain also the variable in terms of types of data collected: continuous, ordinal, categorical

Variables definitions (Operational definitions)

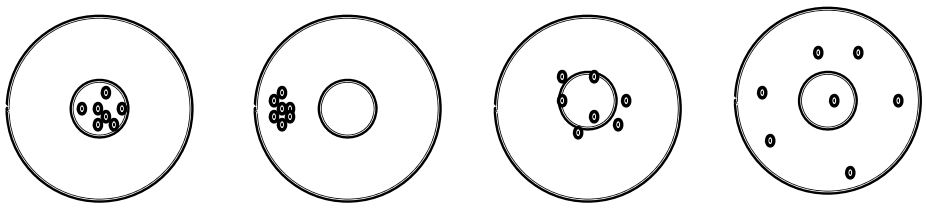
Dependent variable	Definition
Blood Pressure	
Knowledge	
Atherosclerotic changes	
Independent Variable	
Age	
Gender	
Income	
Smoking	
Pesticides exposure	

DATA COLLECTION METHOD

Data collection methods

- Questionnaire
- Clinical assessment
- Blood taking, biological sample management, laboratory methods and analyses
- Environmental sample collection, laboratory methods and analyses

- Data collection method must aim to produce the best **valid results**
- Valid result – high **accuracy** and **precision.**



Data collection method

- Quantitative data can be collected through various methods:
  - Primary Data Collection
    - Surveys
    - Interviews
    - Objective measurement
    - Laboratory analyses
  - Secondary Data Collection
    - Data from Statistic department
    - Data from Patients record

Research instruments

- Questionnaire
- Data collection form
- Clinical/Laboratory equipments
  - weighing scale
  - Stadiometer
  - Spirometer
- Laboratory machines
  - UV spectrophometer
  - ELISA plate reader
  - GCMS
- Laboratory methods

Questionnaire

- Always necessary in any clinical and epidemiology studies
- The crucial part of the questionnaire is on demographic characteristic
- Questionnaire able to measure the subjective matter into objective measurement.
  - Stress
  - Depression
  - Attitudes
  - Knowledge
  - Satisfaction
  - Compliance

Questionnaire

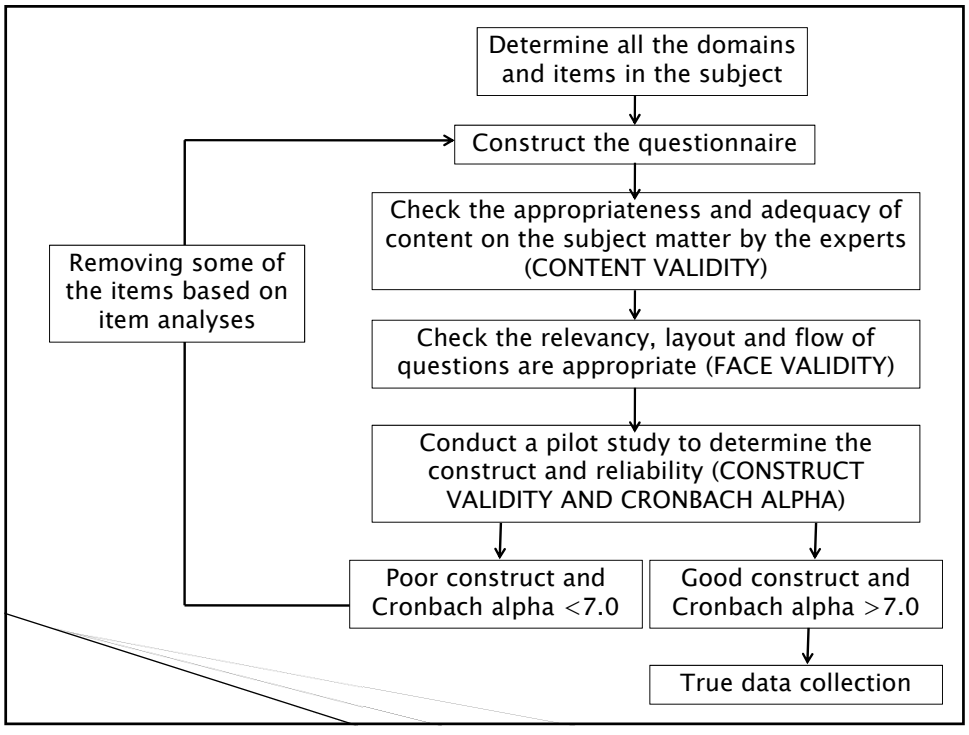
- **Valid** questionnaire:
  - Able to measure what it is suppose to measure and able to achieve the objectives of the research
- Example of validated questionnaire:
  - DASS questionnaire in many languages (English, Malay)
  - SF-36 quality of life questionnaire
- If it is in English, changing into different language will jeopardize its validity

Questionnaire

- **Reproducible** questionnaire
  - Ability of the questionnaire to elicit the same response over time
- **Standardized questionnaire**
  - Questionnaire that already determine and controls of its validity, reproducibility and biases in a specific circumstances and population.

Questionnaire

- Constructing new questionnaire: validation process
  - Determine all the domains and items
  - Content validity
  - Face validity
  - Construct validity



Questionnaire

- Method of delivering questionnaire
  - Self- administrated
  - Assisted-questionnaire / guided – questionnaire (individual/group)
  - Interviewed – questionnaire

Direct measurements from laboratory equipments

- Is the equipment available in the department?
- If it is not available, can you borrow it from other departments? What is the procedure?
- Is the equipment being used by other researchers?
- Is it in good condition?
- When is the last date the equipments was **calibrated** ?
- What are the disposable items required?
- How to run the equipment?
- Is **training** required? Where can I get the training?
- What are the rules in using the equipments to ensure valid results ?
- How to **interpret the data**? How to choose only valid results to be included in the study?

Laboratory analayses

- Automated or manual laboratory analyses
- Accuracy:
  - Run standard samples/calibrators
  - Construct standard curve
  - Ensure the calibrators within ±2SD
- Precision /Repeatability/ Reproducibility
  - Intra-assay coefficient variation
  - Inter-assay coefficient variation

Laboratory analayses

- Intra-assay
  - Intra-assay % CVs should be less than 10.
- Inter-assay
  - Inter-assay % CVs of less than 15 are generally accep table.
- CVs should be calculated from the calculated concentrations rather than the raw optical den sities.
- These scores reflect the performance of the assay in the hands of the user

In this example cortisol concentrations are measured in duplicate for 40 samples. The % CV for each sample is calculated by finding the standard deviation of results 1 and 2, dividing that by the duplicate mean, and multiplying by 100. The average of the individual CVs is reported as the intra-assay CV.

Sample	Result 1	Result 2	Duplicate Mean (µg/dL)	Standard Dev	% CV
1	0.132	0.128	0.130	0.002	1.7
2	0.351	0.361	0.356	0.007	2.0
3	0.282	0.306	0.294	0.017	5.6
4	0.237	0.218	0.227	0.013	5.9
5	0.178	0.215	0.197	0.026	13.5
6	0.119	0.128	0.123	0.006	5.2
7	0.136	0.125	0.131	0.007	5.7
8	0.155	0.150	0.152	0.004	2.6
9	0.145	0.152	0.149	0.005	3.4
37	0.059	0.040	0.050	0.013	27.1
38	0.061	0.059	0.060	0.001	2.4
39	0.201	0.175	0.188	0.018	9.8
40	0.149	0.155	0.152	0.004	2.8

Intra-assay CV (n=40) = average % CV = 5.2

Source: www. salimetrics.com

In this example the same high and low cortisol controls are run in quadruplicate on ten different plates to monitor plate-to-plate variation. The plate means for high and low are calculated and then used to calculate the overall mean, standard deviation, and % CV. Overall % CV = SD of plate means ÷ mean of plate means x 100. The average of the high and low % CV is reported as the inter-assay CV.

Control	Result 1	Result 2	Plate	Plate Mean (µg/dL)
high	1.199	1.132	1	1.090
high	1.071	0.957		
high	1.101	0.973	2	0.998
high	1.007	0.911		
high	1.093	0.942	3	0.976
high	0.918	0.905	9	0.922
high	0.973	0.892		
high	0.986	0.923	10	0.941
high	0.950	0.904		

Control	Result 1	Result 2	Plate	Plate Mean (µg/dL)
low	0.113	0.115	1	0.105
low	0.108	0.083		
low	0.112	0.089	2	0.097
low	0.101	0.085		
low	0.129	0.106	3	0.111
low	0.108	0.104	9	0.106
low	0.110	0.104		
low	0.113	0.107	10	0.103
low	0.099	0.095		

High	
Mean of Means	1.005
Std Dev of Means	0.051
% CV of Means	5.1

Low	
Mean of Means	0.104
Std Dev of Means	0.006
% CV of Means	6.3

Inter-assay CV (n=10) = average of high and low control CV = 5.1 + 6.3 / 2 = 5.7%

THANK YOU