Data management
&
Statistical analysis

Sasivimol Rattanasiri, Ph.D

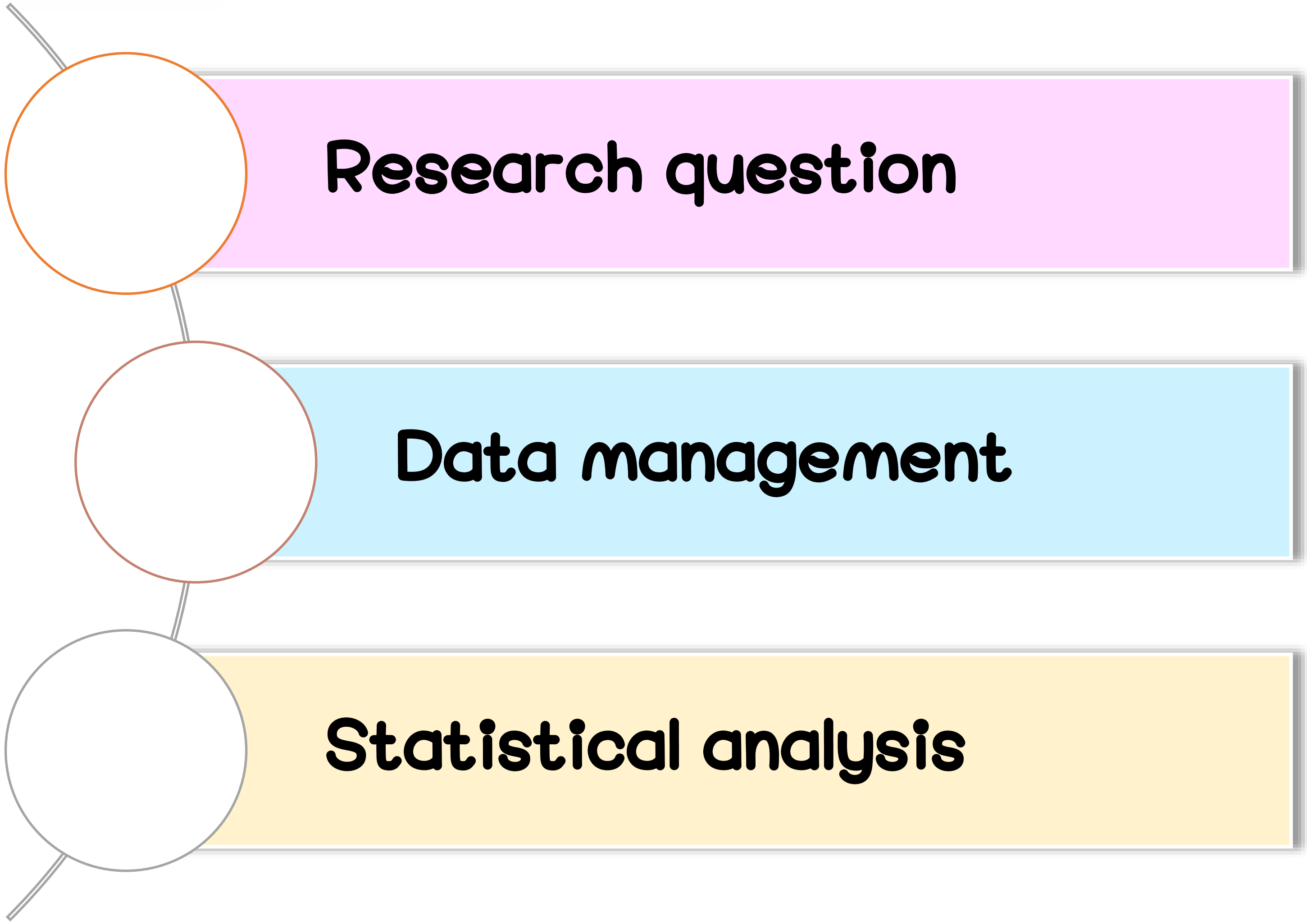
Department of Clinical Epidemiology and Biostatistics

Ramathibodi Hospital, Mahidol University

E-mail: sasivimol.rat@mahidol.edu

www.ceb-rama.org

Scope



Research question

Data management

Statistical analysis

Scope

Research question

Data management

Statistical analysis

Understand basic questions…

What are objectives of research?

What is type of study design?

What variables will be involved?

How variables will be measured?

How often variables will be collected?

Understand basic questions…

What are objectives of research?

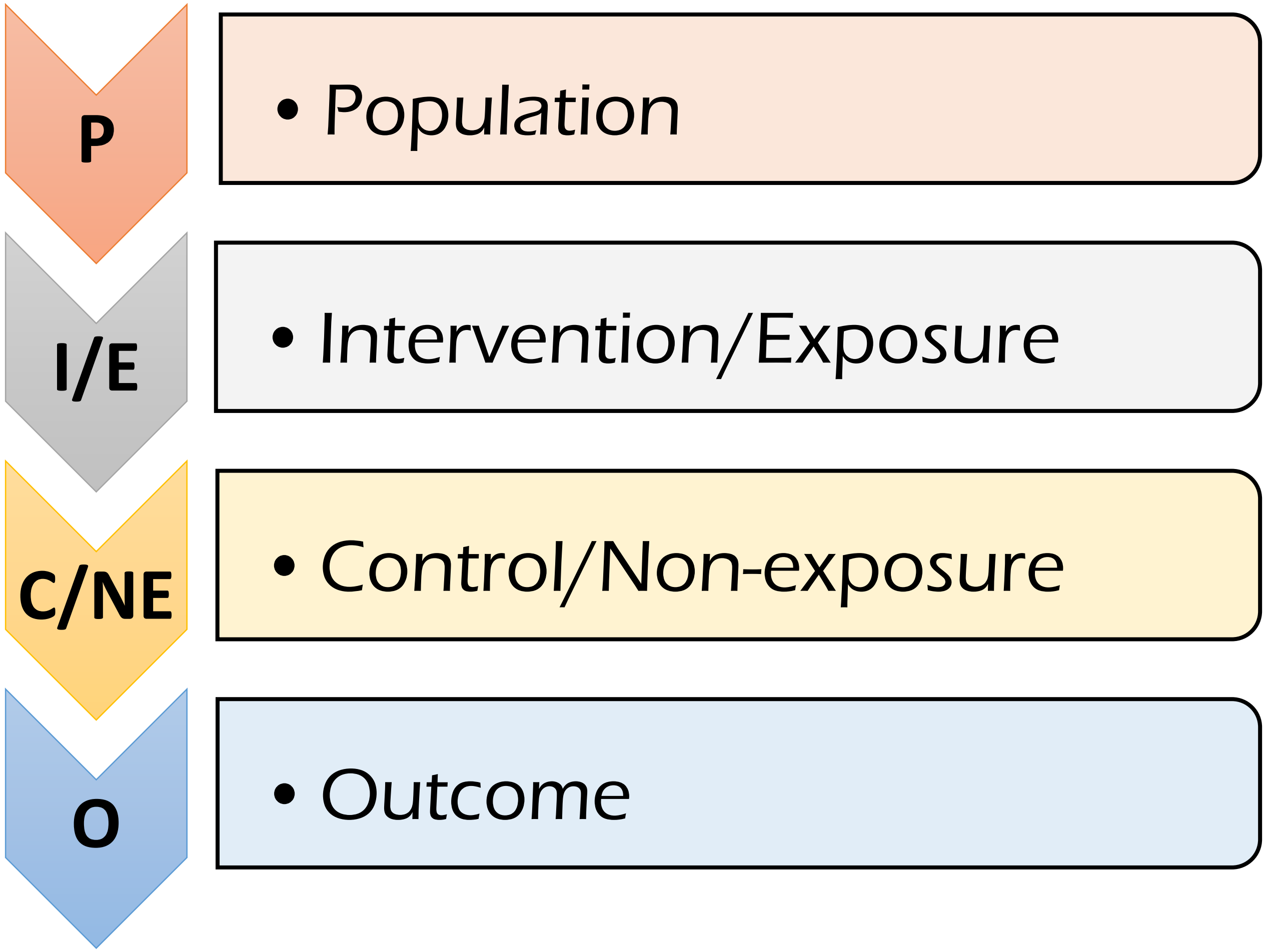
What is type of study design?

What variables will be involved?

How variables will be measured?

How often variables will be collected?

PICOS



P
- Population

I/E
- Intervention/Exposure

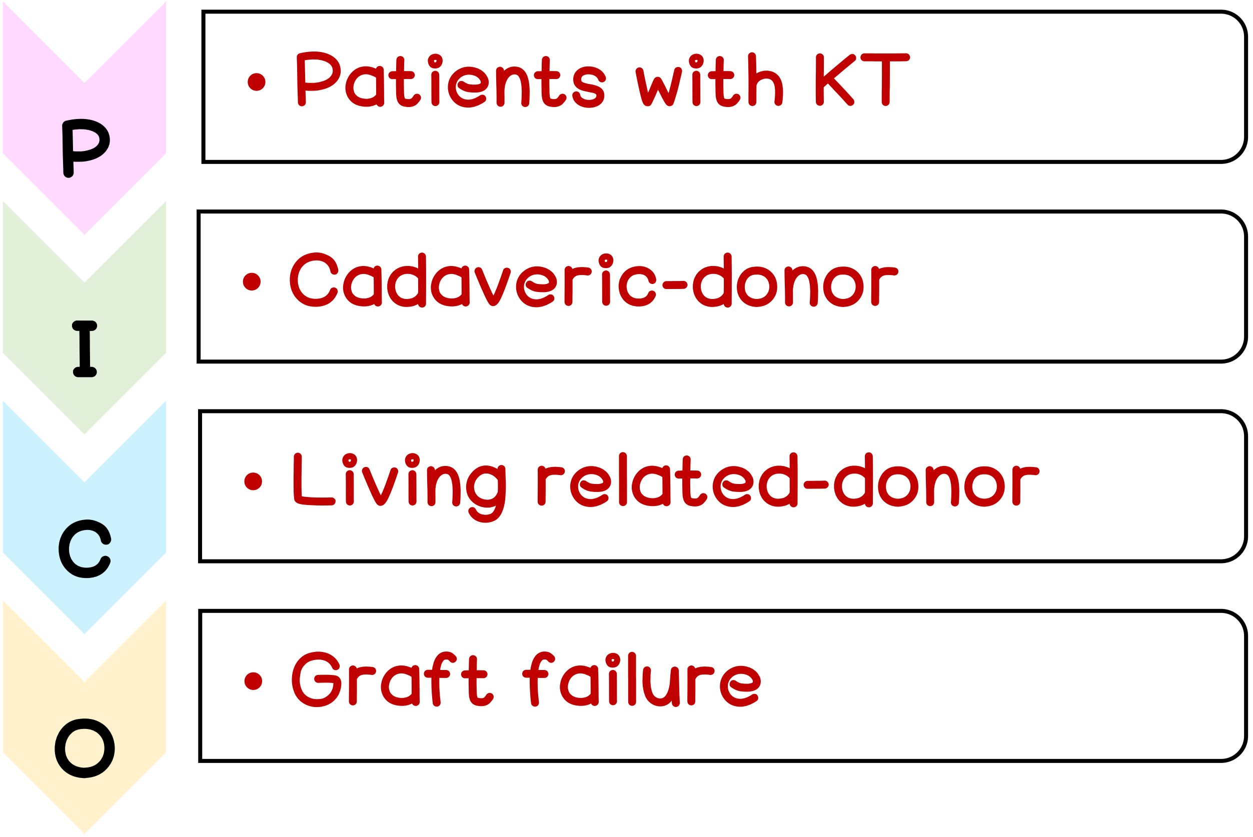
C/NE
- Control/Non-exposure

O
- Outcome

Research question

A retrospective cohort study of kidney transplantation (KT) patients was conducted to assess the association between types of donors and risk of graft failure

PICOS



P
- Patients with KT

I
- Cadaveric-donor

C
- Living related-donor

O
- Graft failure

Understand basic questions…

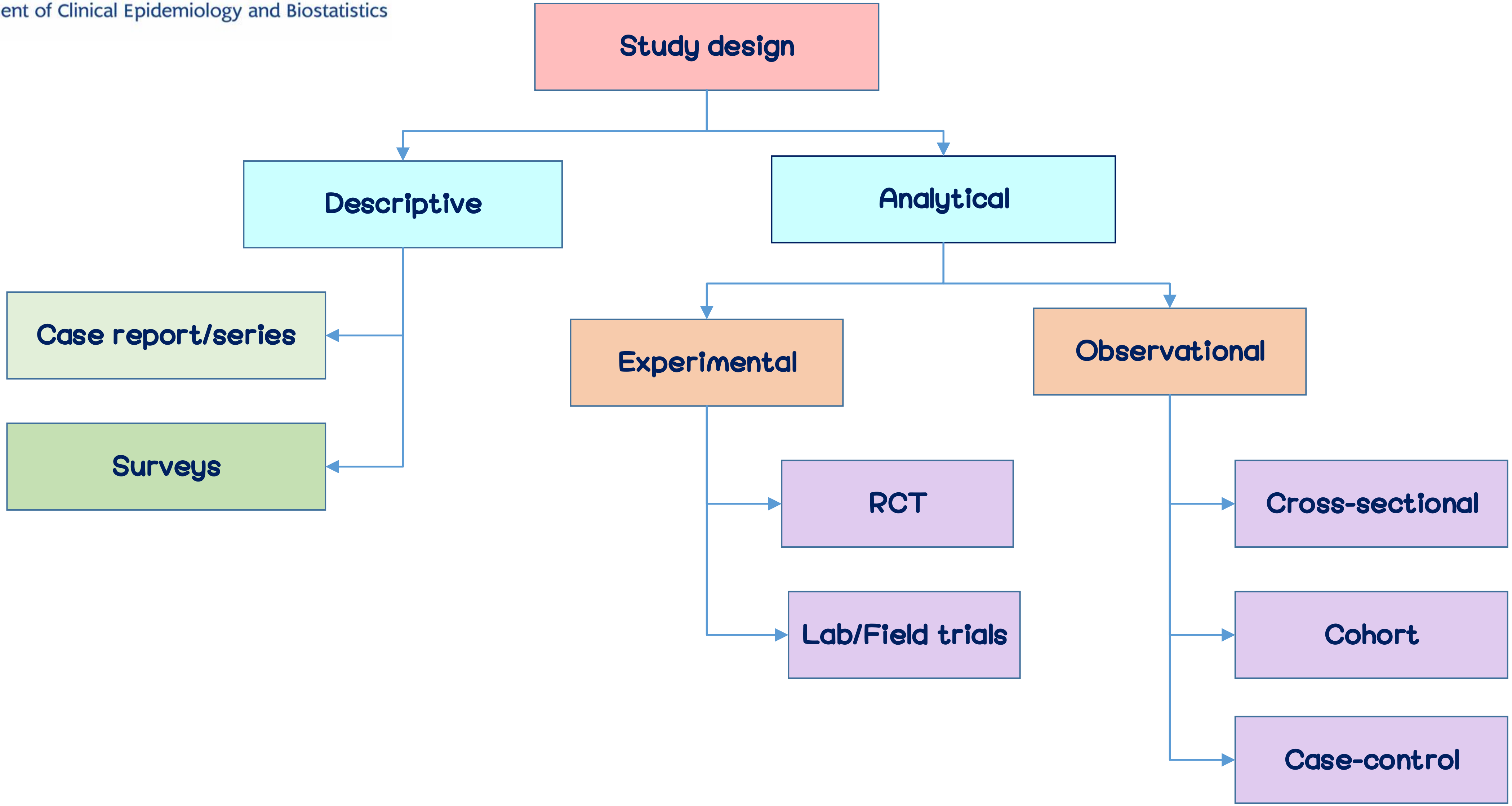
What are objectives of research?

What is type of study design?

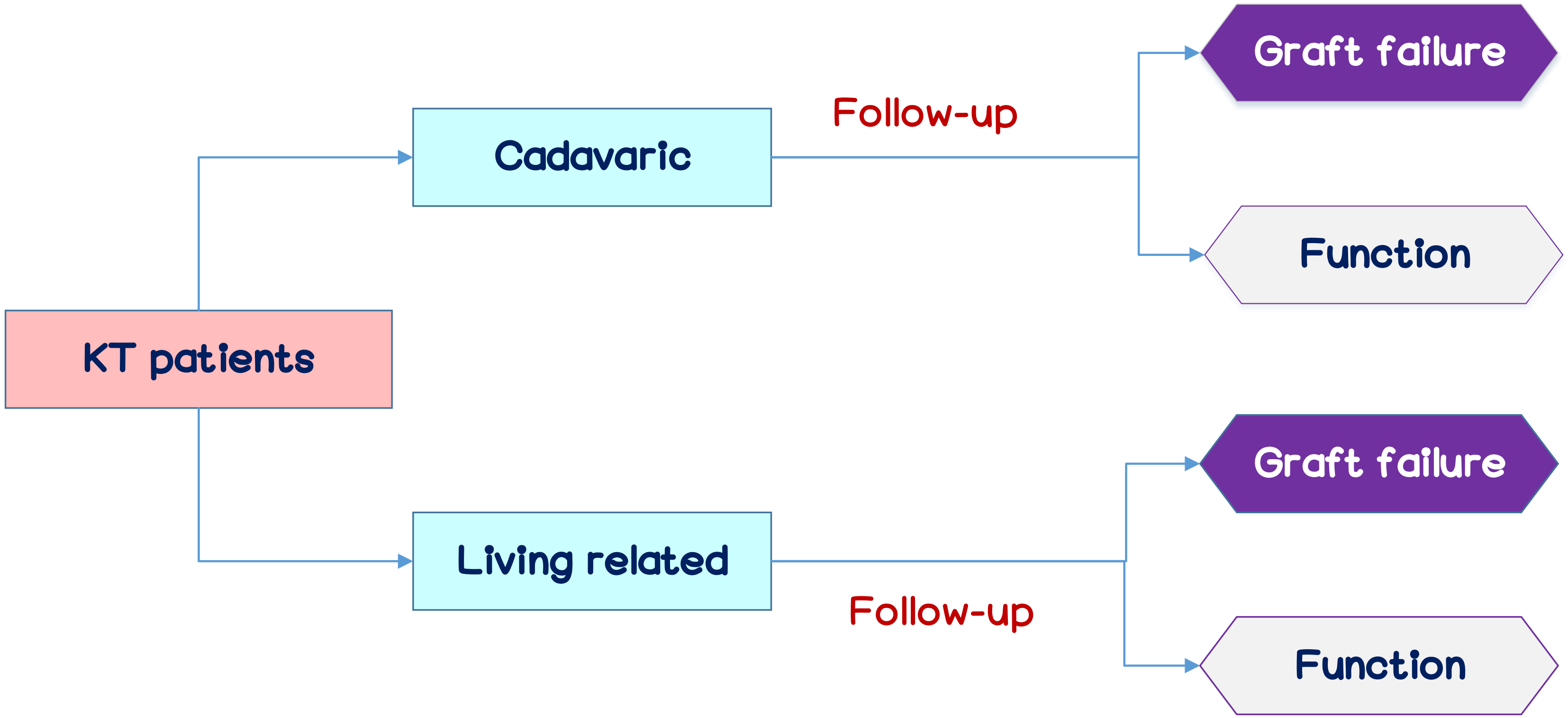
What variables will be involved?

How variables will be measured?

How often variables will be collected?


Study design
Descriptive
Analytical
Case report/series
Surveys
Experimental
Observational
RCT
Lab/Field trials
Cross-sectional
Cohort
Case-control

A retrospective cohort study of KT patients



KT patients

Cadavaric

Follow-up

Graft failure

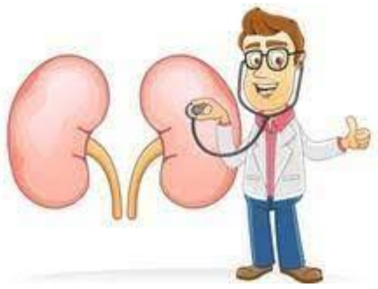
Function

Living related

Follow-up

Graft failure

Function



Understand basic questions…

What are objectives of research?

What is type of study design?

What variables will be involved?

How variables will be measured?

How often variables will be collected?

Factors associated with graft failure

BMI

Gender

Age

Underlying disease
- DM - CVD
- HT - CAD

Pre-transplant illness

Type of donor

Understand basic questions…

What are objectives of research?

What is type of study design?

What variables will be involved?

How variables will be measured?

How often variables will be collected?

Factors associated with graft failure

1. Date of birth	□□ / □□ / □□□□ (DD/MM/YYYY)	
2. Gender	□1. Male	□ 2. Female
3. Types of donor	□1. CDKT	□ 2. LRKT
4. Weight	□□□ kg.	
5. Height	□□□ cm.	

Study factor

Factors associated with graft failure

1. Date of visit	□□ / □□ / □□□□ (DD/MM/YYYY)
2. Graft status	□1. failure　　□ 2. function
3. Date of failure	□□ / □□ / □□□□ (DD/MM/YYYY)
4. Serum creatinine	□□□ mg/dL
5. Serum albumin	□□□ g/dL

Outcome

Understand basic questions…

What are objectives of research?

What is type of study design?

What variables will be involved?

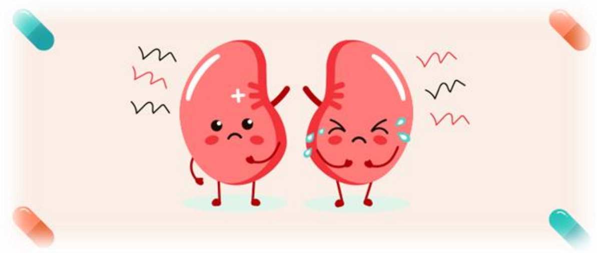
How variables will be measured?

How often variables will be collected?

How often variable will be collected…

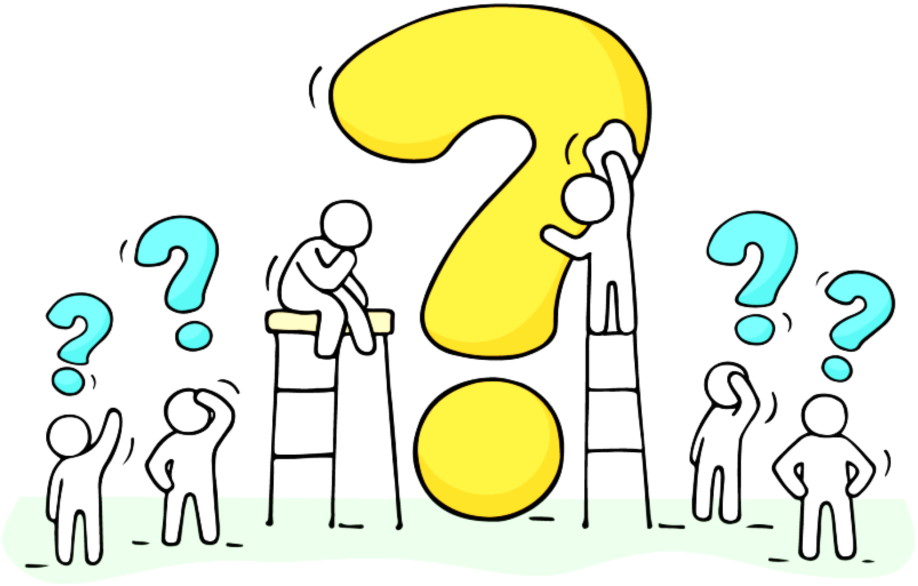
Types of donor

→ collected at the enrollment period

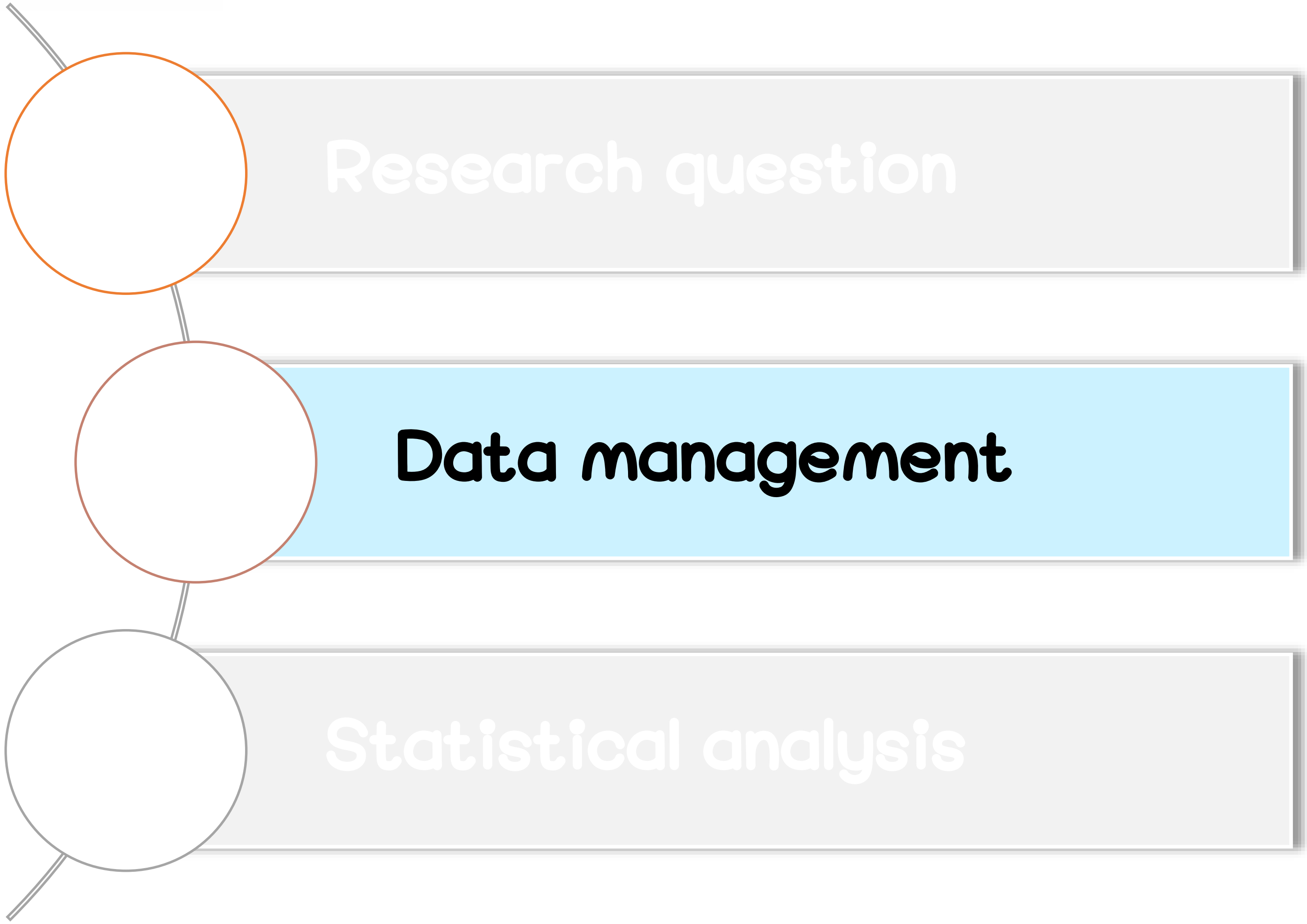


Graft status

→ collected every 6 months after KT



Scope

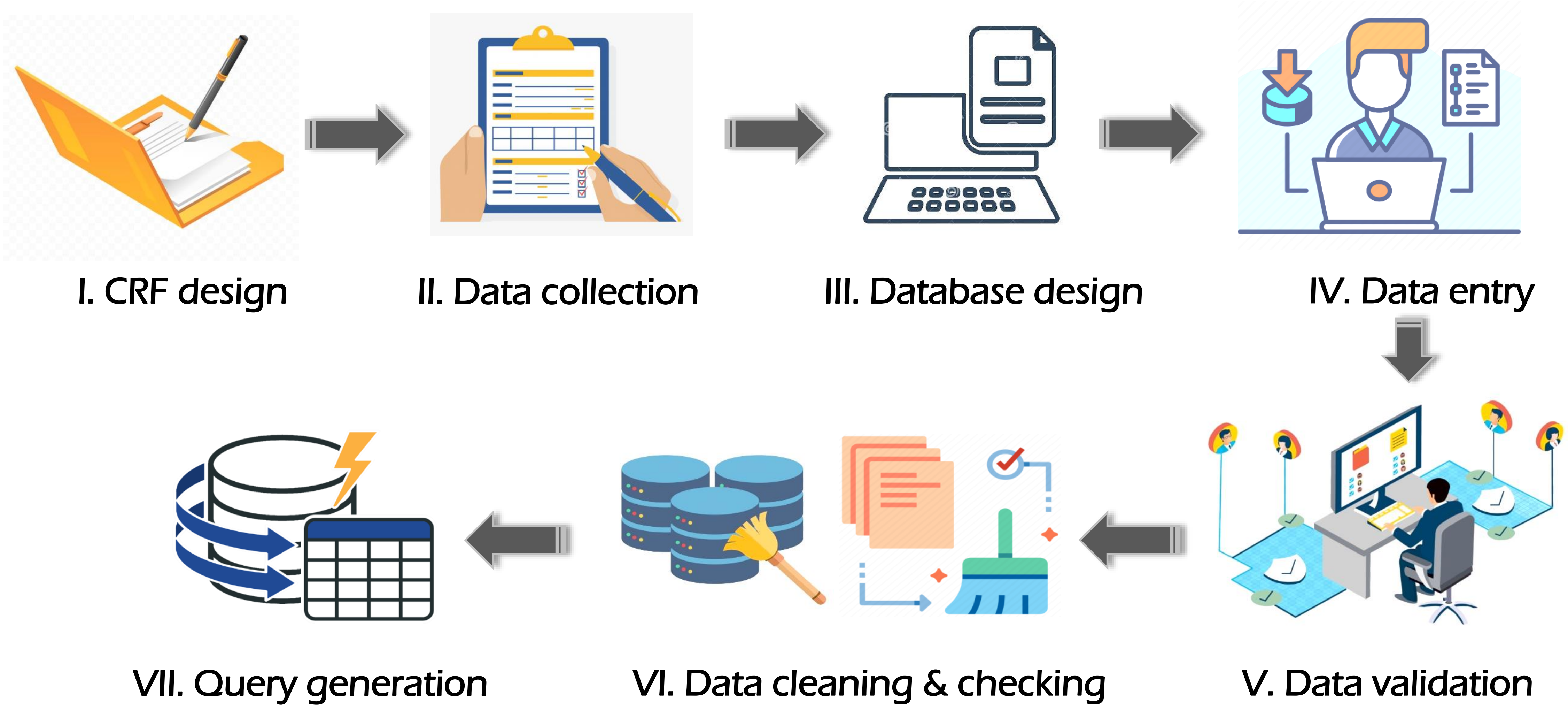


Research question

Data management

Statistical analysis

Data management



I. CRF design

II. Data collection

III. Database design

IV. Data entry

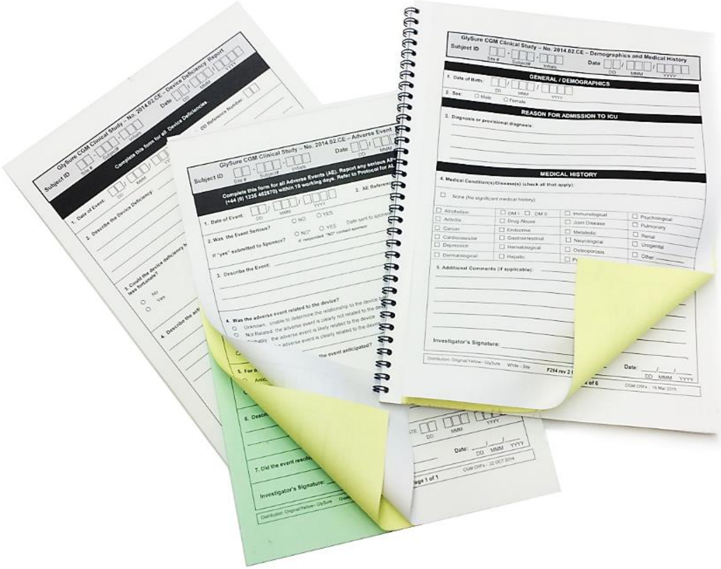
V. Data validation

VI. Data cleaning & checking

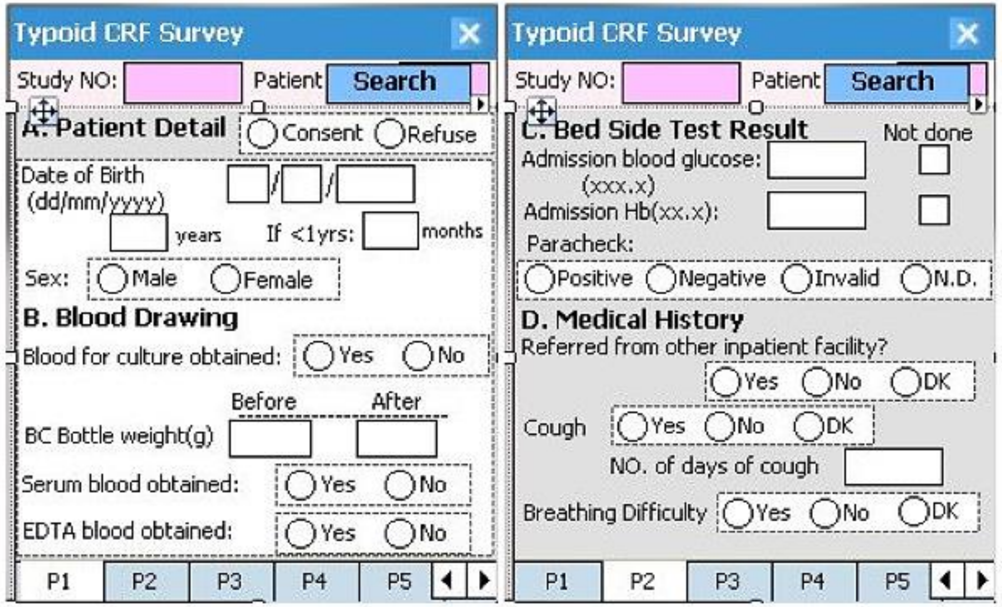
VII. Query generation

Case Record Form (CRF)

- a paper or electronic form designed to collect all of data , **which specifies by study protocol**



Paper CRF



Typoid CRF Survey

Study NO: Patient Search

A. Patient Detail ○Consent ○Refuse

Date of Birth (dd/mm/yyyy) / /

years If <1yrs: months

Sex: ○Male ○Female

B. Blood Drawing

Blood for culture obtained: ○Yes ○No

	Before	After
BC Bottle weight(g)		

Serum blood obtained: ○Yes ○No

EDTA blood obtained: ○Yes ○No

P1 P2 P3 P4 P5

Typoid CRF Survey

Study NO: Patient Search

C. Bed Side Test Result Not done

Admission blood glucose: (xxx.x) □

Admission Hb(xx.x): □

Paracheck:

○Positive ○Negative ○Invalid ○N.D.

D. Medical History

Referred from other inpatient facility?

○Yes ○No ○DK

Cough ○Yes ○No ○DK

NO. of days of cough

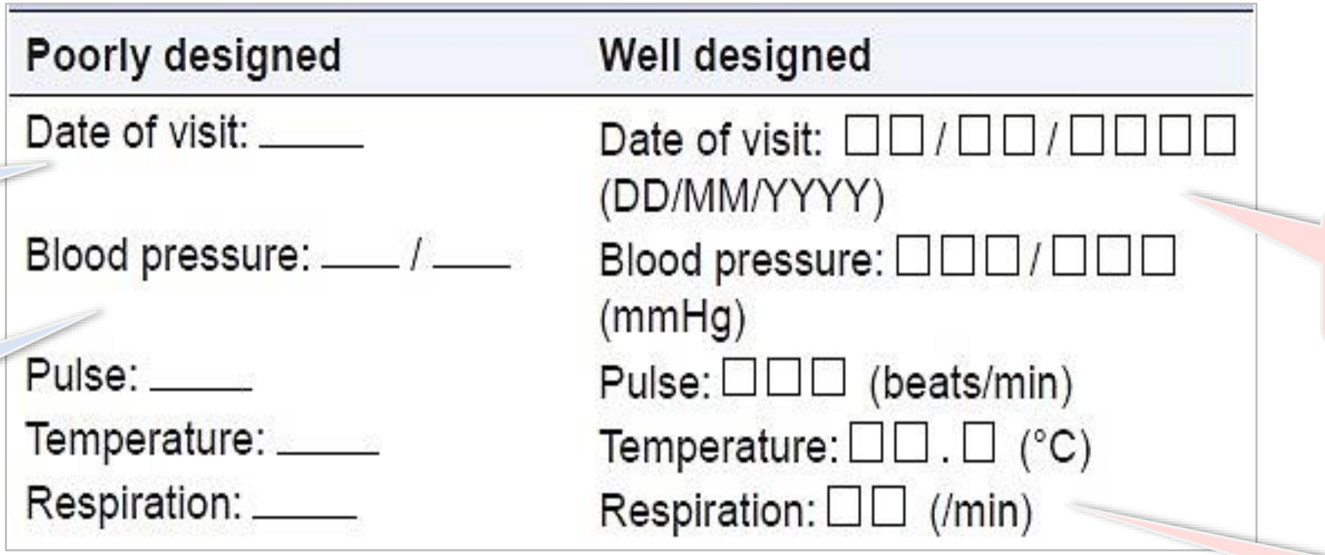
Breathing Difficulty ○Yes ○No ○DK

P1 P2 P3 P4 P5

Electronic CRF

Poorly designed CRF

Well designed CRF



Poorly designed	Well designed
Date of visit: ____	Date of visit: □□/□□/□□□□ (DD/MM/YYYY)
Blood pressure: ___/___	Blood pressure: □□□/□□□ (mmHg)
Pulse: ____	Pulse: □□□ (beats/min)
Temperature: ____	Temperature: □□.□ (°C)
Respiration: ____	Respiration: □□ (/min)

No boxes to hold answers

Unit of measurement did not display on CRF

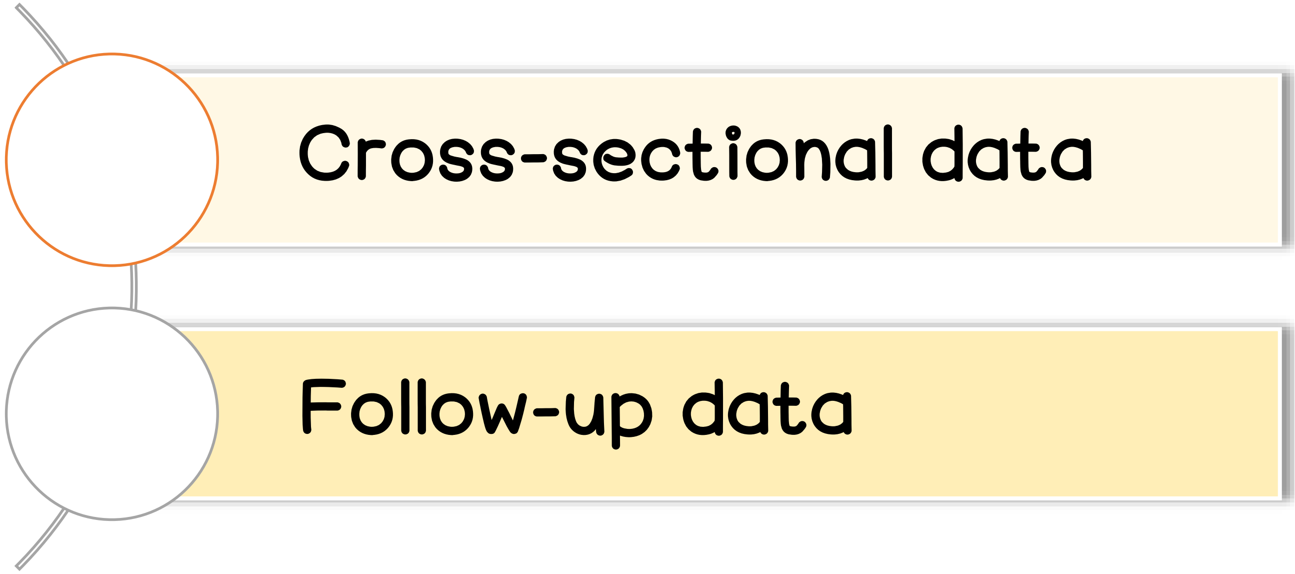
Provide boxes to hold answers

Units and decimal points should be displayed

Objective of CRF design…

Preserve and maintain	Quality and integrity of data
Gather	Complete and accurate data
Avoid	Duplication of data
Facilitate	Transcription of data from sources documents onto CRF

How to prepare data in Excel…



Cross-sectional data

Follow-up data

Cross-sectional data

Subject	Age	Sex	Group	Response
1	32	Female	Treatment	No
2	45	Female	Control	No
3	23	Male	Control	Yes
4	38	Female	Treatment	No
5	36	Male	Control	Yes
6	29	Male	Control	Yes
7	43	Male	Treatment	Yes
8	39	Female	Control	No
9	51	Male	Treatment	Yes
10	42	Female	Treatment	No

Variable name

- Not exceed than 10 characters
- Not contain space
- Not begin with number

Cross-sectional data

Subject	Age	Sex	Group	Response
1	32	Female	Treatment	No
2	45	Female	Control	No
3	23	Male	Control	Yes
4	38	Female	Treatment	No
5	36	Male	Control	Yes
6	29	Male	Control	Yes
7	43	Male	Treatment	Yes
8	39	Female	Control	No
9	51	Male	Treatment	Yes
10	42	Female	Treatment	No

Types of Data

- Only numerical data
- Set special for missing data

Cross-sectional data

Subject	Sex	Group	Response
1	Female	Treatment	No
2	Female	Control	No
3	Male	Control	Yes
4	Female	Treatment	No
5	Male	Control	Yes

1. Male
2. Female

1. Treatment
2. Control

1. Yes
2. No

Subject	Sex	Group	Response
1	2	1	2
2	2	2	2
3	1	2	1
4	2	1	2
5	1	2	1

Cross-sectional data

Subject	DM	HT	CVD	Malignant
1	1	2	1	2
2	1	2	2	2
3	2	1	2	1
4	1	2	1	2
5	2	1	2	1
6	2	1	2	1
7	1	1	1	1
8	1	2	2	2
9	2	1	1	1
10	1	2	1	2

Use consistency code

1. Yes
2. No

Cross-sectional data

Subject	TAC	Dose_TAC	MMF	Dose_MMF
1	1	0.5	1	180
2	1	1.5	2	0
3	2	0	2	0
4	1	3.5	1	360
5	2	0	2	0
6	2	0	2	0
7	1	2.5	1	180
8	1	3	2	0
9	2	0	1	540
10	1	2.5	1	360

Dose format

- Specify unit for data entry (mg/day)
- Enter "0" for not receive treatment

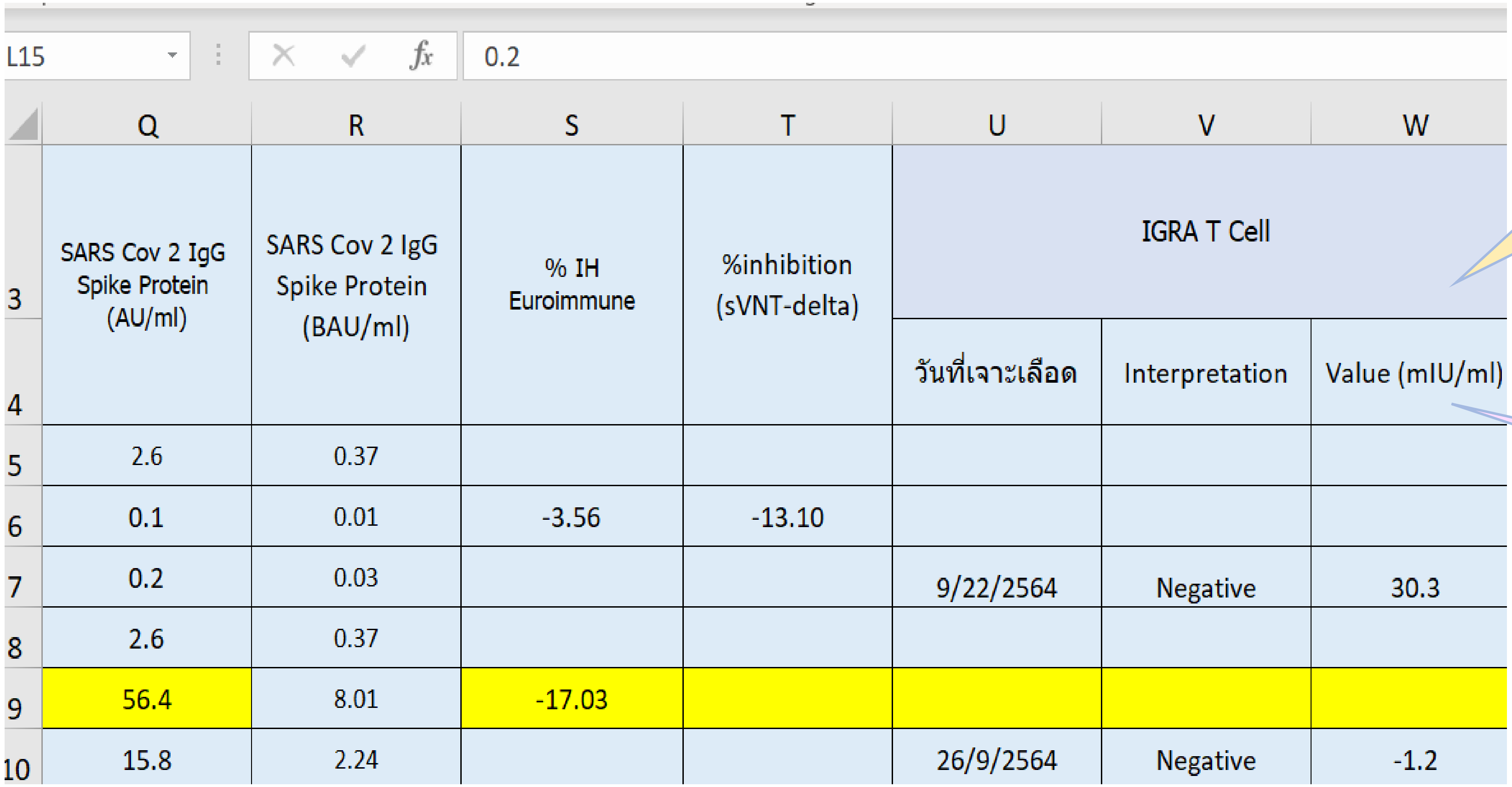
Follow-up data: Long format

Subject	Visit	Date visit	SBP	DBP	Response
1	1	12/08/2000	90	65	2
1	2	05/09/2000	95	60	2
1	3	11/12/2000	90	65	1
2	1	16/03/2011	120	80	2
2	2	09/07/2011	125	85	1
3	1	06/12/2012	100	85	2
3	2	08/06/2013	105	90	2
3	3	10/09/2013	110	90	2
4	1	23/04/2008	150	95	2
4	2	19/11/2008	155	98	1

Follow-up data: Wide format

Subject	date1	SBP1	DBP1	resp1	date2	SBP2	DBP2	resp2	date…	SBP…
1										
2										
3										
4										
5										

Inappropriate data format



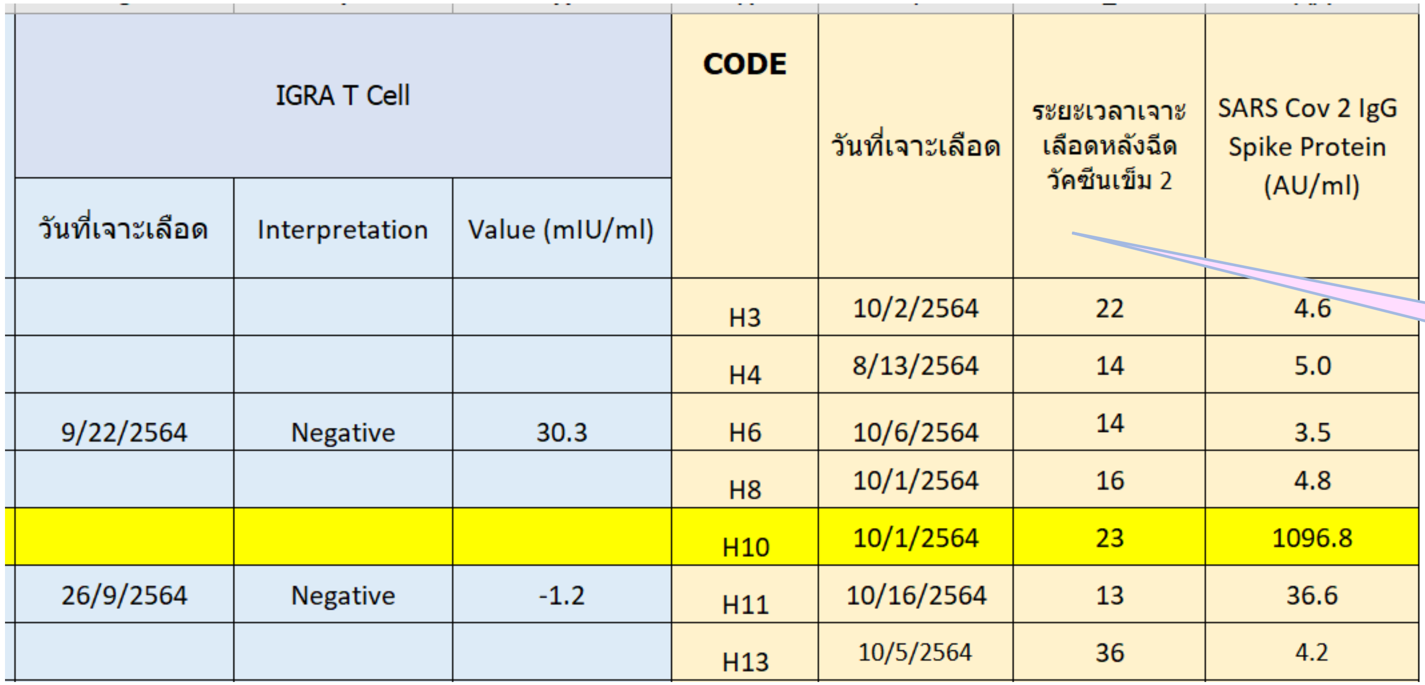
L15 fx 0.2

	Q	R	S	T	U	V	W
3	SARS Cov 2 IgG Spike Protein (AU/ml)	SARS Cov 2 IgG Spike Protein (BAU/ml)	% IH Euroimmune	%inhibition (sVNT-delta)	IGRA T Cell		
4					วันที่เจาะเลือด	Interpretation	Value (mIU/ml)
5	2.6	0.37					
6	0.1	0.01	-3.56	-13.10			
7	0.2	0.03			9/22/2564	Negative	30.3
8	2.6	0.37					
9	56.4	8.01	-17.03				
10	15.8	2.24			26/9/2564	Negative	-1.2

Not appropriate format for statistical analysis

Incorrect variable name

Inappropriate data format

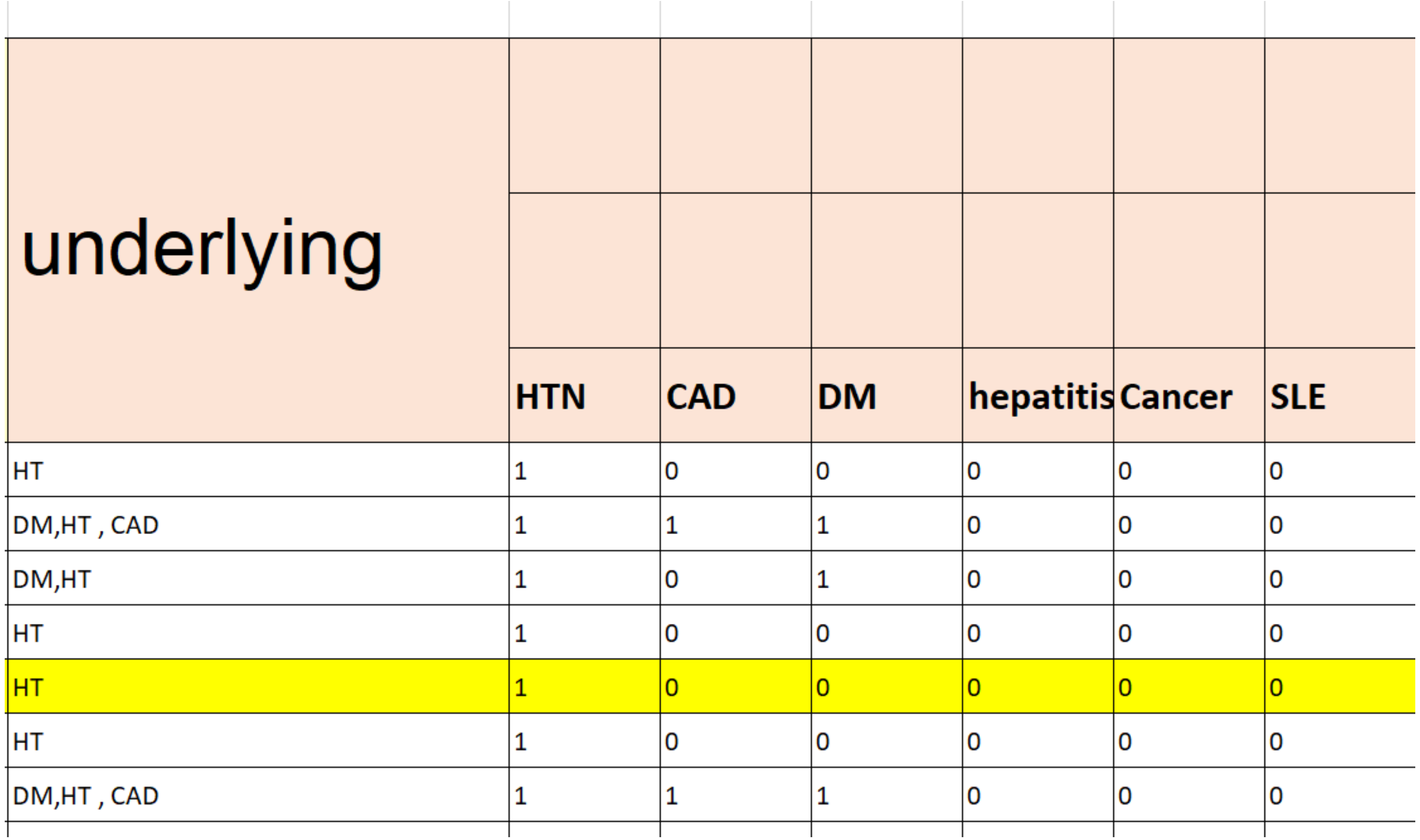


IGRA T Cell			CODE	วันที่เจาะเลือด	ระยะเวลาเจาะเลือดหลังฉีดวัคซีนเข็ม 2	SARS Cov 2 IgG Spike Protein (AU/ml)
วันที่เจาะเลือด	Interpretation	Value (mIU/ml)				
			H3	10/2/2564	22	4.6
			H4	8/13/2564	14	5.0
9/22/2564	Negative	30.3	H6	10/6/2564	14	3.5
			H8	10/1/2564	16	4.8
			H10	10/1/2564	23	1096.8
26/9/2564	Negative	-1.2	H11	10/16/2564	13	36.6
			H13	10/5/2564	36	4.2

Incorrect variable name

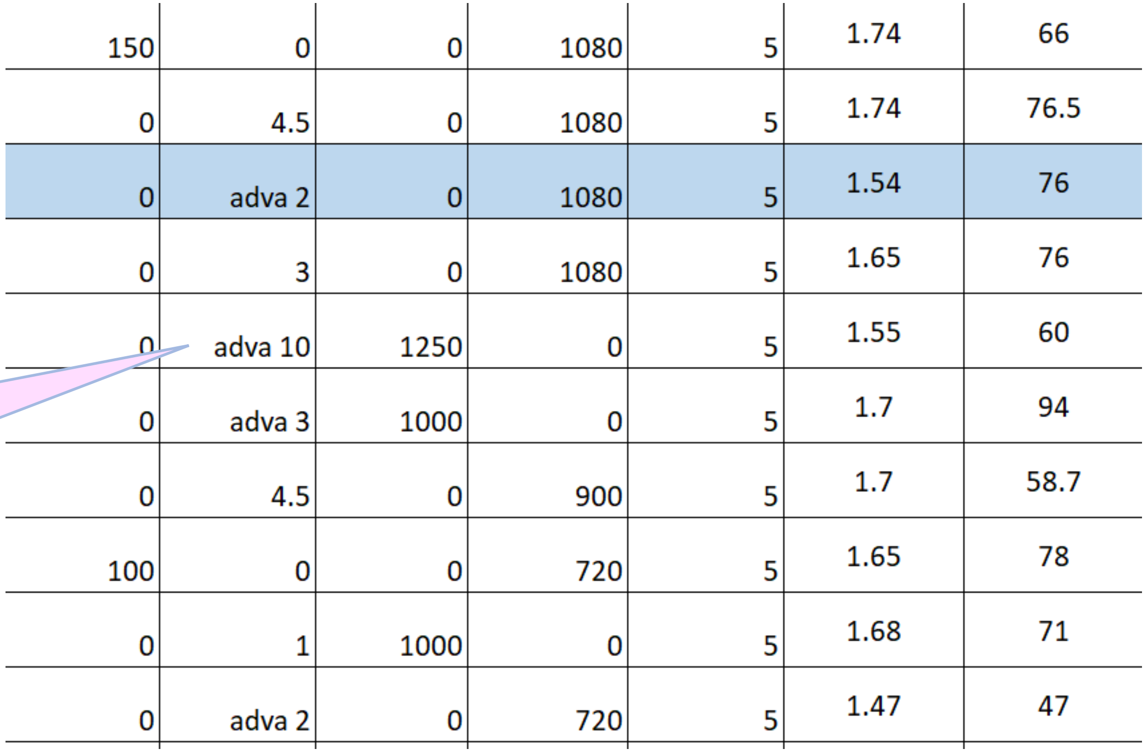
Inappropriate data format

Not appropriate format for underlying disease



underlying	HTN	CAD	DM	hepatitis	Cancer	SLE
HT	1	0	0	0	0	0
DM,HT , CAD	1	1	1	0	0	0
DM,HT	1	0	1	0	0	0
HT	1	0	0	0	0	0
HT	1	0	0	0	0	0
HT	1	0	0	0	0	0
DM,HT , CAD	1	1	1	0	0	0

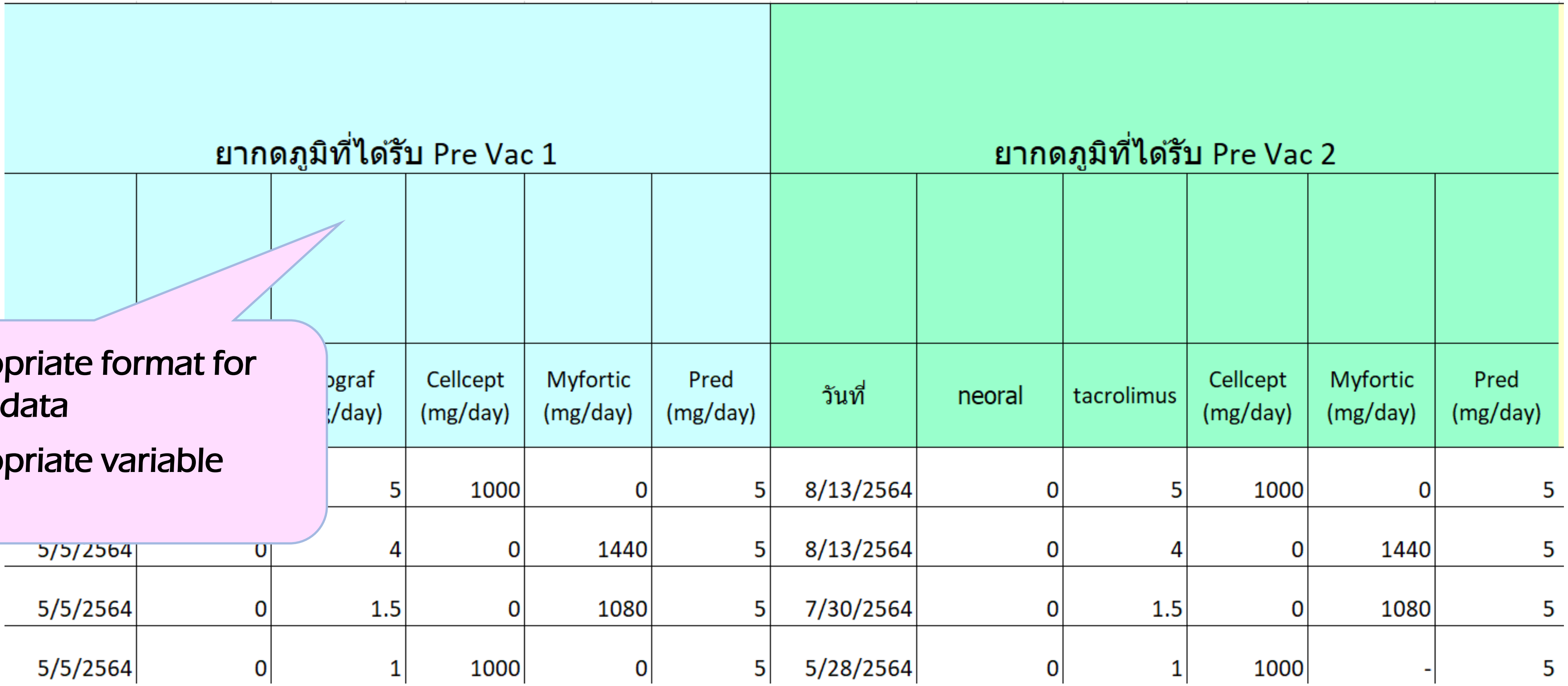
Inappropriate data format



150	0	0	1080	5	1.74	66
0	4.5	0	1080	5	1.74	76.5
0	adva 2	0	1080	5	1.54	76
0	3	0	1080	5	1.65	76
0	adva 10	1250	0	5	1.55	60
0	adva 3	1000	0	5	1.7	94
0	4.5	0	900	5	1.7	58.7
100	0	0	720	5	1.65	78
0	1	1000	0	5	1.68	71
0	adva 2	0	720	5	1.47	47

Not appropriate format for dose

Inappropriate data format



ยากดภูมิที่ได้รับ Pre Vac 1						ยากดภูมิที่ได้รับ Pre Vac 2					
		graf /day)	Cellcept (mg/day)	Myfortic (mg/day)	Pred (mg/day)	วันที่	neoral	tacrolimus	Cellcept (mg/day)	Myfortic (mg/day)	Pred (mg/day)
		5	1000	0	5	8/13/2564	0	5	1000	0	5
5/5/2564	0	4	0	1440	5	8/13/2564	0	4	0	1440	5
5/5/2564	0	1.5	0	1080	5	7/30/2564	0	1.5	0	1080	5
5/5/2564	0	1	1000	0	5	5/28/2564	0	1	1000	-	5

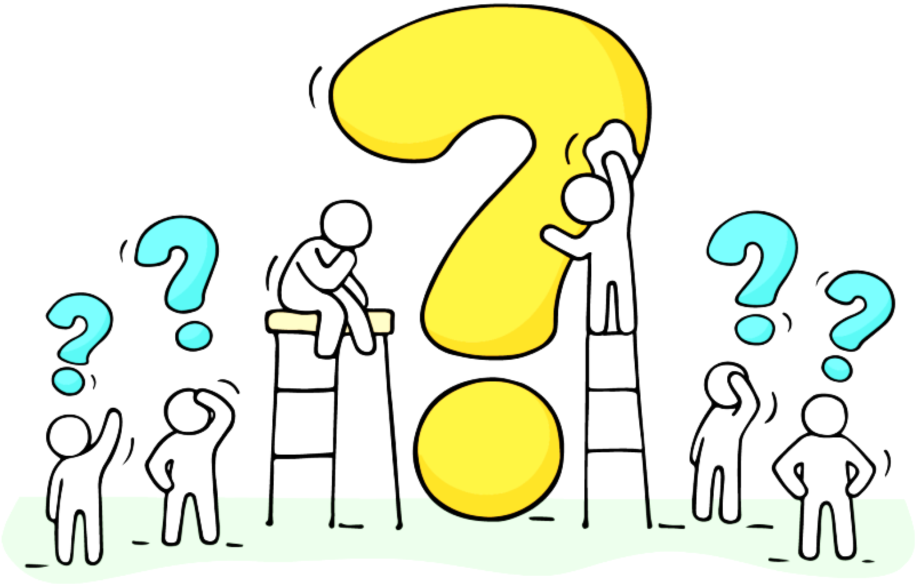
- Not appropriate format for follow-up data
- Not appropriate variable name

Appropriate long format for follow-up data

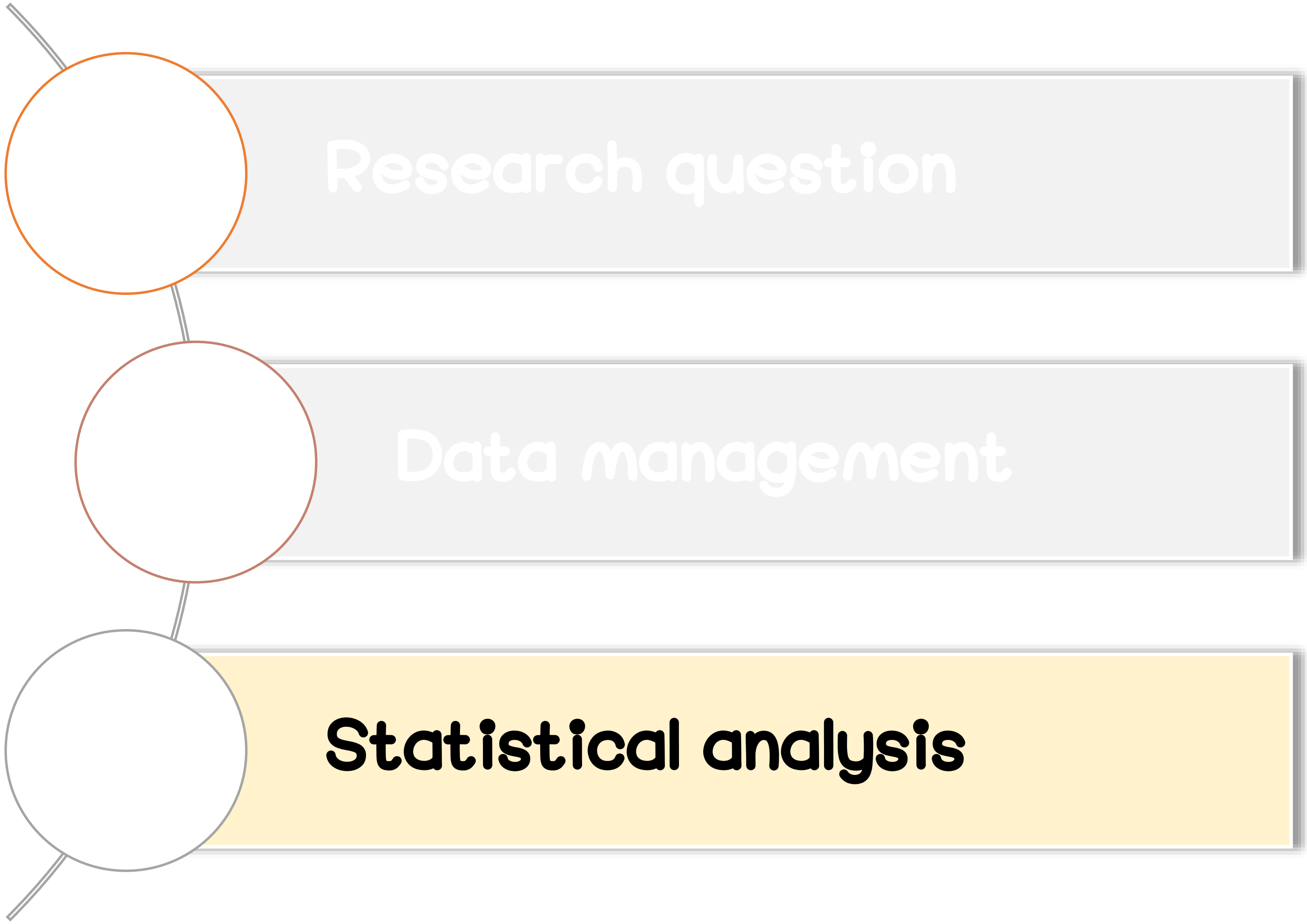
Subject	vaccine	neoral	prograf	cellcept	myfortic	pred
1	1	0	5	1000	0	5
1	2	0	4	1500	180	5
2	1	0	1.5	1000	360	5
2	2	100	1	0	720	5
3	1	0	0	1250	0	5
3	2	150	4	0	180	5
4	1	0	2	1000	0	5
4	2	0	0	1500	720	5
5	1	100	3.5	0	360	5
5	2	150	0	0	180	5

Appropriate wide format for follow-up data

Subject	neoral1	prograf1	cellcept1	myfortic1	neoral2	prograf2	cellcept2	myfortic2
1	0	5	1000	0	100	5	1000	0
2	0	4	1500	180	150	4	1500	180
3	0	1.5	1000	360	0	1.5	1000	360
4	100	1	0	720	0	1	0	720
5	0	0	1250	0	100	0	1250	0
6	150	4	0	180	0	4	0	180
7	0	2	1000	0	150	2	1000	0
8	0	0	1500	720	100	0	1500	720
9	100	3.5	0	360	0	3.5	0	360
10	150	0	0	180	150	0	0	180



Scope



Research question

Data management

Statistical analysis

I. Basic concept for statistics

- Types of data
- Descriptive statistics
- Inferential statistics

II. Hypothesis testing

- Categorical outcome
- Continuous outcome

I. Basic concept for statistics

- Types of data
- Descriptive statistics
- Inferential statistics

II. Hypothesis testing

- Categorical outcome
- Continuous outcome

Types of data

Categorical data

- Nominal data
- Ordinal data

Numerical data

- Discrete data
- Continuous data

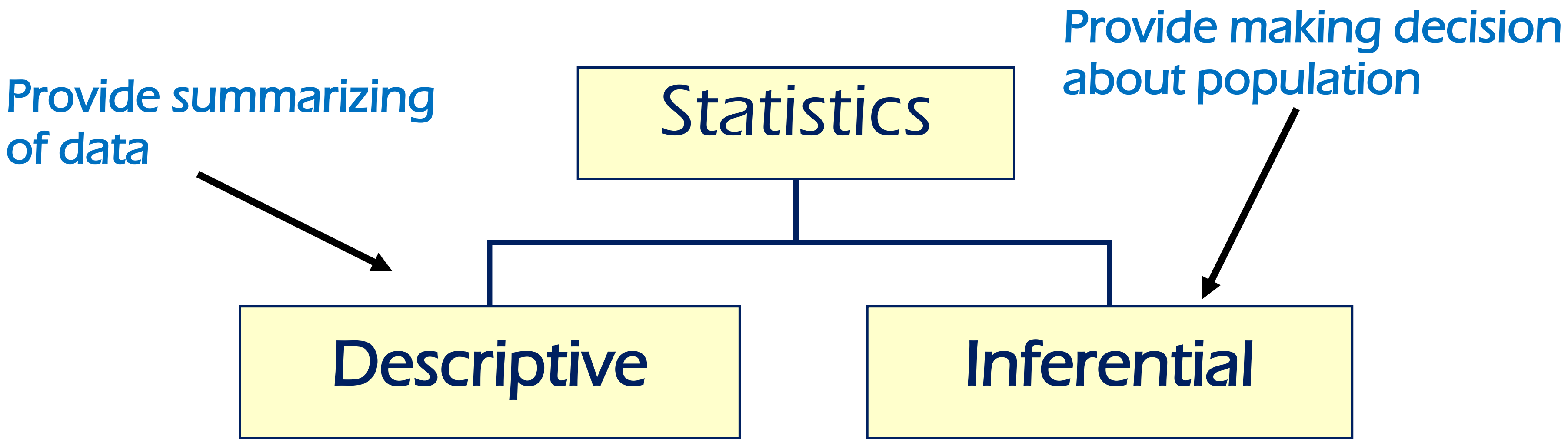
Categorical data

Nominal data	Sex: male/female –>dichotomous data
	Blood group: A/B/AB/O
Ordinal data	Degree of injury: mild/moderate/severe
	Stage of cancer: I/II/III/IV

Numerical data

Discrete data	Length of hospital stay
	Number of heart beats per minute
Continuous data	Cholesterol level (mg/dL)
	Fasting blood sugar (mg/dL)

Types of statistics



Statistics

Provide summarizing of data

Descriptive

Provide making decision about population

Inferential

- Mean, median
- SD, Range

- t-test
- Chi-square test
- Analysis of variance
- Regression analysis

I. Basic concept for statistics

- Types of data
- Descriptive statistics
- Inferential statistics

II. Hypothesis testing

- Categorical outcome
- Continuous outcome

Summarizing: Categorical data

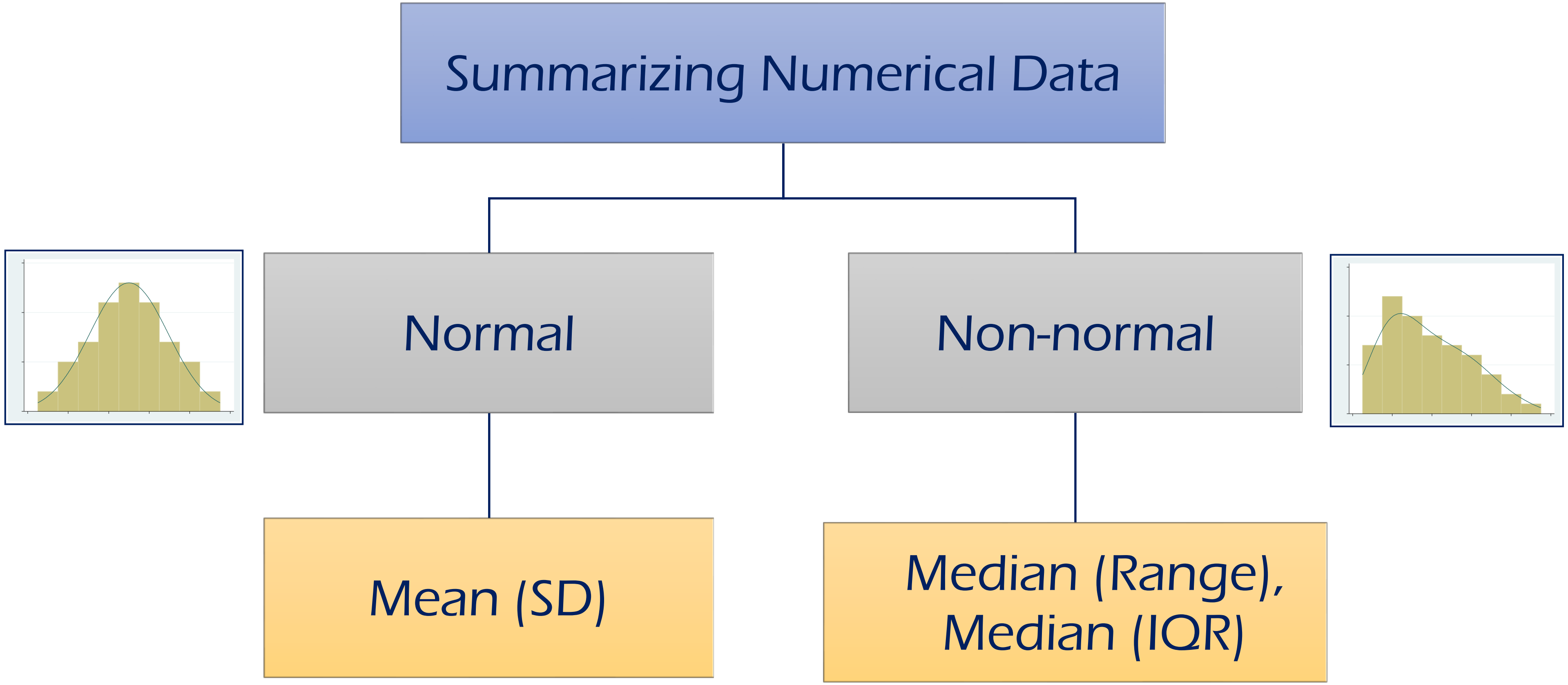
Sex	Frequency	Percentage
Male	56	80
Female	14	20
Total	70	100

Nominal data

Ordinal data

Stages of cancers	Frequency	Percentage
I	120	15
II	320	40
III	160	20
IV	200	25
Total	800	100

Summarizing: Numerical data



Summarizing Numerical Data

Normal

Non-normal

Mean (SD)

Median (Range),
Median (IQR)

Summarizing: Numerical data

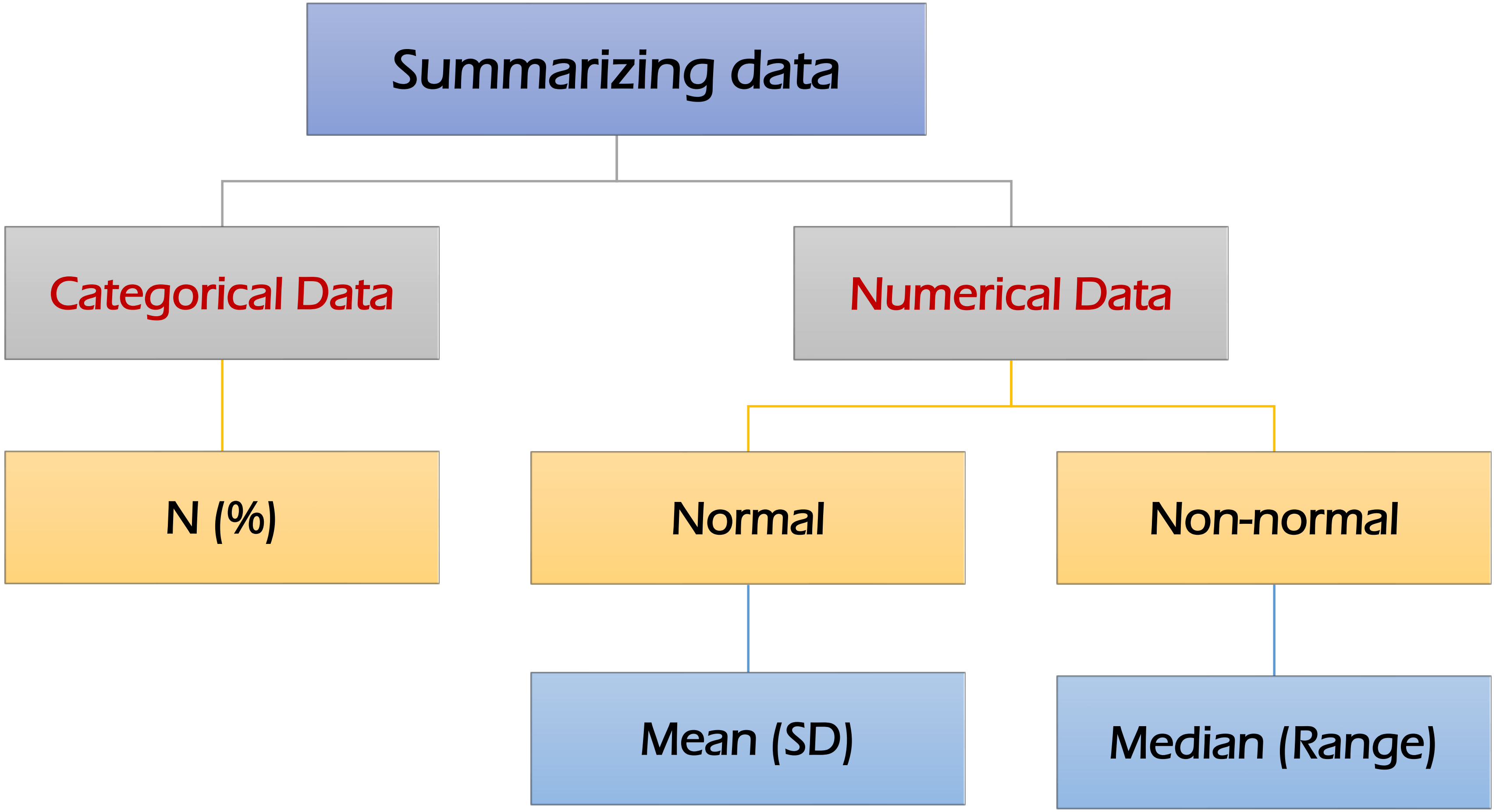
	Mean	SD
Age (year)	49.6	14.3
Weight (cm)	95.6	21.7
Height (cm)	161.5	9.2

Normal distribution

	Mean	SD
CD4 count	62.4	~~74.4~~
~~CA score~~	177.7	352.9
	Median	Range
CD4 count	30.5	1,358
CA score	51.0	1,4879

Non-normal distribution

Summarizing data



Dummy table for descriptive data

Characteristics
Gender; n (%)
Male
Female
Age; years; mean (sd)
Age; n (%)
<30 years
≥30 years
Body weight; kg; mean (sd)
Diabetes; n (%)
Yes
No

questions

I. Basic concept for statistics

- Types of data
- Descriptive statistics
- Inferential statistics

II. Hypothesis testing

- Categorical outcome
- Continuous outcome

Inferential statistics

Parameter estimation

- Point estimate
- Range estimate

Hypothesis testing

- Single population
- Two population
- More than two pop.

Inferential statistics

Parameter estimation

- Point estimate
- Range estimate

Hypothesis testing

- Single population
- Two population
- More than two pop.

Parameter estimation

- Estimation of **mean age** of patients who had breast cancer in Thailand
- Estimation **prevalence of chronic kidney disease** in Thai population

Continuous data

Point estimate	The mean age of 750 patients with DM was 54.42 years
Range estimate	95% CI of mean age range from 43.65 years to 62.34 years

Categorical data

Point estimate	42 in 350 subjects had hypertension, prevalence of hypertension was 0.12
Range estimate	95% CI of the prevalence of hypertension was from 0.09 to 0.15

Recommendation

- Point estimate should be reported with their confidence intervals to indicate their precision
- Prevalence of HT was 12% with 95% CI: 9-15%

Inferential statistics

Parameter estimation

- Point estimate
- Range estimate

Hypothesis testing

- Single population
- Two population
- More than two pop.

Hypothesis testing

Continuous outcome

Test if the means of BMD between postmenopausal women who received and did not receive calcium supplements differ.

Dichotomous outcome

Assess the association between traditional medicine used and osteoporotic hip fracture.

Types of errors

Ho: $\mu_{BMD(calcium+)} = \mu_{BMD(calcium-)}$

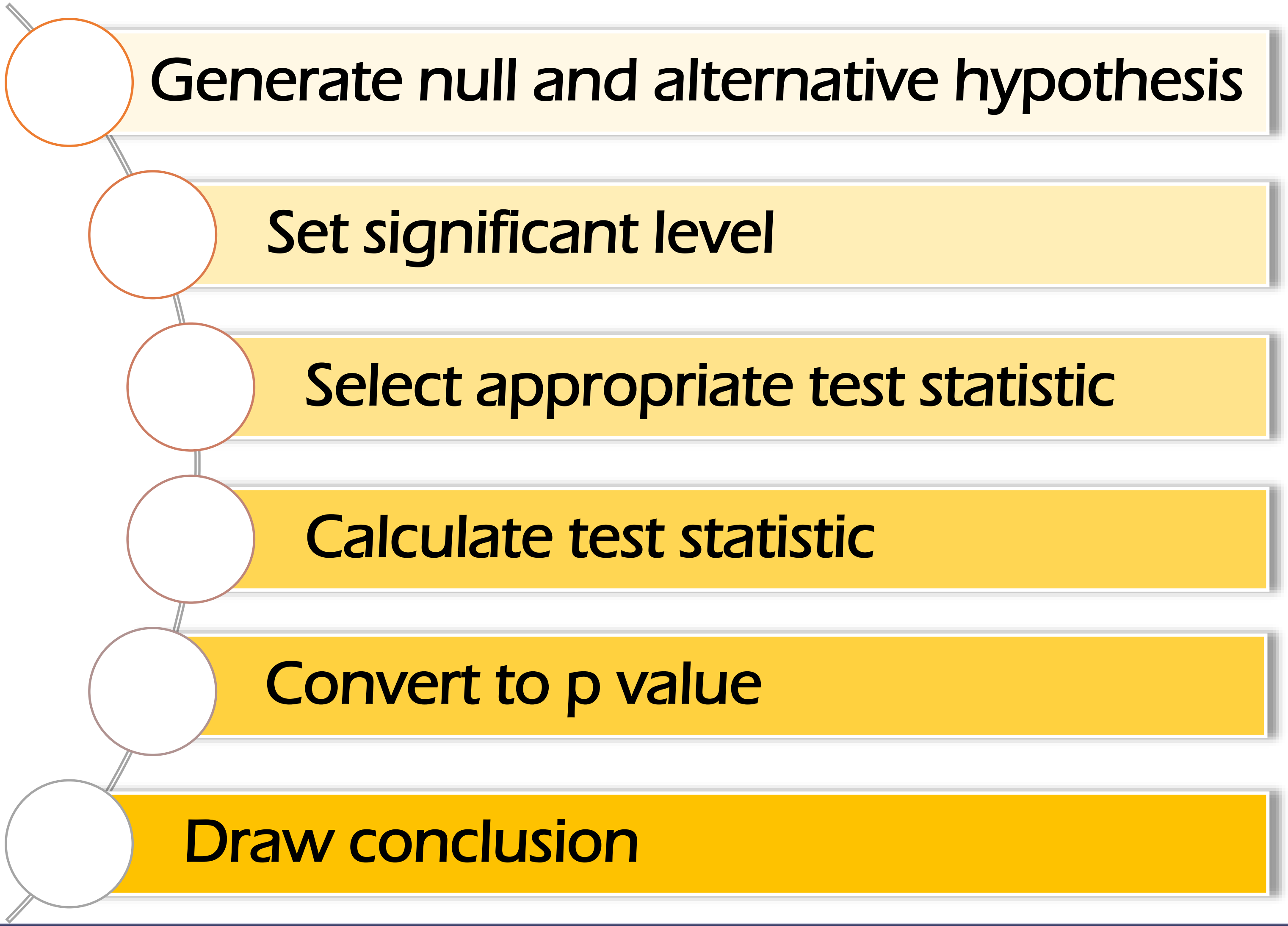
Ha: $\mu_{BMD(calcium+)} \neq \mu_{BMD(calcium-)}$

Known

Unknown

Statistical Decision	In Population	
Based on Sample	H_0 is true	H_0 is false
Reject H_0	a *(Type I error)*	1- b *(Power of test)*
Do not reject H_0	1- a *(Confidence)*	b *(Type II error)*

Steps of hypothesis testing



- Generate null and alternative hypothesis
- Set significant level
- Select appropriate test statistic
- Calculate test statistic
- Convert to p value
- Draw conclusion

Hypothesis testing

Null hypothesis

Ho: $\mu_{BMD(calcium+)} = \mu_{BMD(calcium-)}$

Alternative hypothesis

Ha: $\mu_{BMD(calcium+)} \neq \mu_{BMD(calcium-)}$

I. Generate Ho and Ha

II. Set significance level

Significance level	Test statistics	P value	Decision
0.05	0.935	0.358	Fail to reject Ho
0.05	-3.884	< 0.01	Reject Ho

III. Calculate statistic

IV. Calculate P value

V. Draw conclusion

I. Basic concept for statistics

- Types of data
- Descriptive statistics
- Inferential statistics

II. Hypothesis testing

- **Categorical outcome**
- Continuous outcome

Hypothesis testing for categorical data

Tests of association
Independent sample
Paired-sample

Hypothesis testing for categorical data

Tests of association
Independent sample
Paired-sample

Independent sample

- A case-control study was conducted to look at effect of traditional medicine and osteoporotic hip fracture.
- The outcome of interest was osteoporotic hip fracture.
- The exposure of interest was traditional medicine.

2x2 contingency table for independent sample

Hip fracture	Traditional medicine used		
	Yes	No	n
Yes	20	208	228
No	8	216	224

Data layout

Statistical analysis

- The Chi-square test is used to examine association between two categorical variables
- H_0: The proportions of the interested event between two independent groups are not different
- H_0: Two categorical variables are independent

H_0: No association between traditional medicine and hip fracture

Conclusion

- Reject null hypothesis
- There was association between traditional medicine and hip fracture

```
. tab tredmed hip,col exp chi2

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
| column percentage  |
+--------------------+

traditiona |     hip fracture
l medicine |       yes         no |     Total
-----------+----------------------+----------
       yes |        20          8 |        28
           |      14.1       13.9 |      28.0
           |      8.77       3.57 |      6.19
-----------+----------------------+----------
        no |       208        216 |       424
           |     213.9      210.1 |     424.0
           |     91.23      96.43 |     93.81
-----------+----------------------+----------
     Total |       228        224 |       452
           |     228.0      224.0 |     452.0
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   5.2588   Pr = 0.022
```

Statistical analysis

- The Chi-square test is not appropriate if small sample.
- Expected frequency is less than 5 for more than 20% of the total cells
- The Fisher's exact test is an alternative method

Independent with small sample

- A case-control study was conducted to look at effect of receiving HRT on risk of hip fracture.
- The outcome of interest was hip fracture.
- The exposure of interest was HRT.

2x2 contingency table for independent sample

Hip fracture	HRT		
	Yes	No	n
Yes	1 (1.5)	213 (212.5)	214
No	2 (1.5)	214 (214.5)	216

Data layout

H_0: No association between HRT and hip fracture

Conclusion

- Fail to reject null hypothesis
- There was no association between HRT and hip fracture

```
. tab hrt hip,col exp exact

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
| column percentage  |
+--------------------+

           |          hip
       hrt |       yes         no |     Total
-----------+----------------------+----------
       yes |         1          2 |         3
           |       1.5        1.5 |       3.0
           |      0.47       0.93 |      0.70
-----------+----------------------+----------
        no |       213        214 |       427
           |     212.5      214.5 |     427.0
           |     99.53      99.07 |     99.30
-----------+----------------------+----------
     Total |       214        216 |       430
           |     214.0      216.0 |     430.0
           |    100.00     100.00 |    100.00

           Fisher's exact =                 1.000
   1-sided Fisher's exact =                 0.503
```

Dummy table for two groups comparison

Characteristics	Hip fracture	Non-hip fracture	P value
	n (%)	n (%)	
Age, year			
< 60			
≥ 60			
Gender			
Male			
Femal			
Hypertension			
Yes			
No			

Hypothesis testing for categorical data

Tests of association
Independent sample
Paired-sample

Paired sample

- ➲ Comparison of pain relief (yes/no) by two different analgesics in the same subjects.
- ➲ In a matched case-control study, matched case to control patients with BMI, aim to assess the association between HRT and the hip fracture.

2x2 contingency table for paired sample

Case	Control		
	HRT+	HRT-	n
HRT+	102	50	152
HRT-	100	120	220

Data layout

H_0: No association between HRT and hip fracture

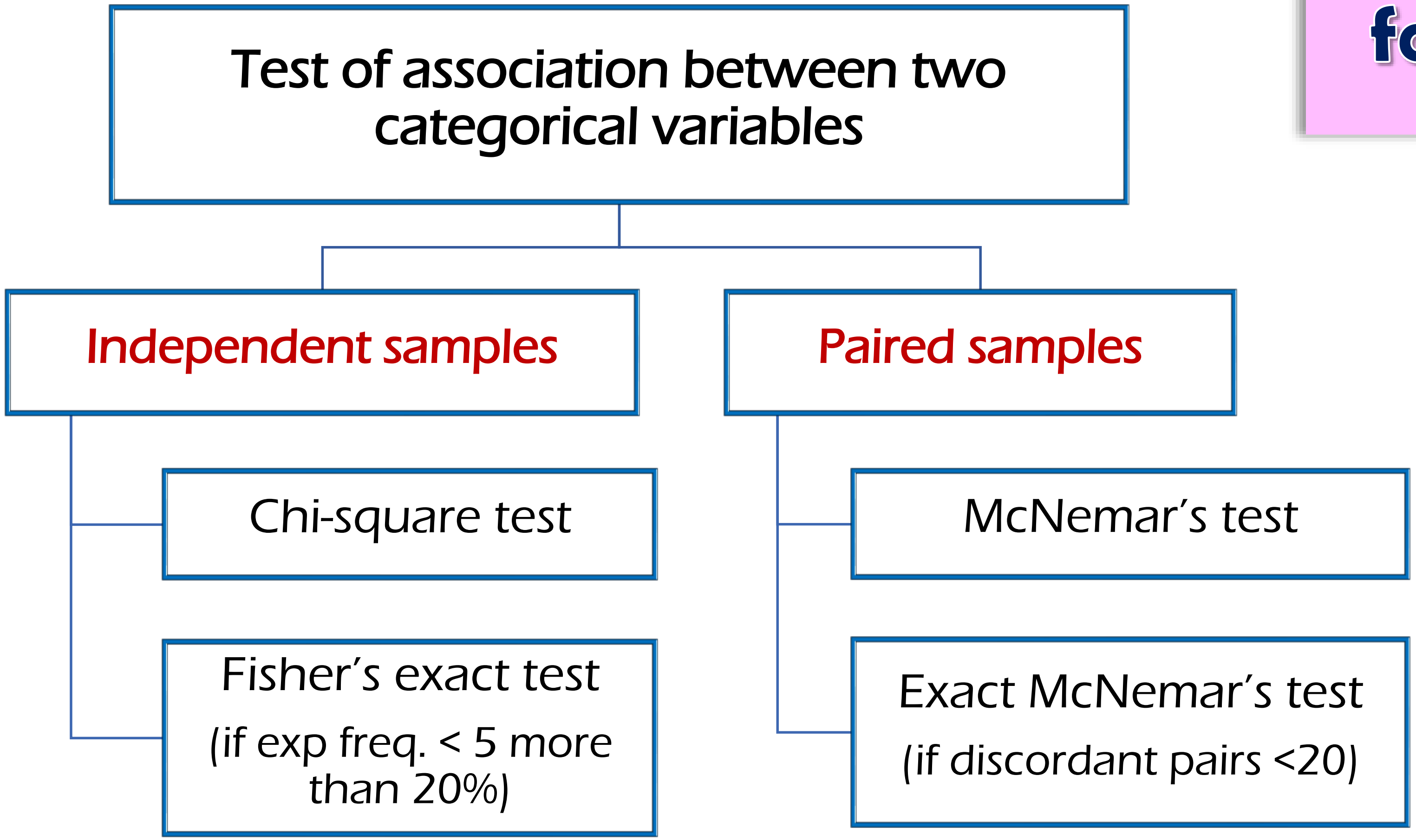
Conclusion

- ➲ Reject null hypothesis
- ➲ There was association between HRT and hip fracture

```
. mcc case control

                 | Controls               |
Cases            |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
         Exposed |       102          50  |        152
       Unexposed |       100         120  |        220
-----------------+------------------------+------------
           Total |       202         170  |        372

McNemar's chi2(1) =     16.67    Prob > chi2 = 0.0000
Exact McNemar significance probability       = 0.0001

Proportion with factor
        Cases       .4086022
        Controls    .5430108     [95% Conf. Interval]
                   ---------     --------------------
        difference -.1344086     -.2001631  -.0686541
        ratio       .7524752       .656141   .8629532
        rel. diff. -.2941176     -.4547495  -.1334858

        odds ratio        .5      .3487202   .7089431   (exact)
```

Note: if number of discordant pairs is less than 20, the Exact McNemar's test is more appropriate

Hypothesis testing for categorical data



Test of association between two categorical variables

Independent samples

Chi-square test

Fisher's exact test

(if exp freq. < 5 more than 20%)

Paired samples

McNemar's test

Exact McNemar's test

(if discordant pairs <20)

I. Basic concept for statistics

- Types of data
- Descriptive statistics
- Inferential statistics

II. Hypothesis testing

- Categorical outcome
- **Continuous outcome**

Hypothesis testing for continuous data

Single group

Two groups

- Independent sample
- Paired sample

Three groups or more

Hypothesis testing for continuous data

Single group

Two groups

- Independent sample
- Paired sample

Three groups or more

Independent sample

- ➲ Comparison of systolic blood pressure between men and women.
- ➲ Comparison of cholesterol level between patients with and without chronic kidney disease.

Statistical test for two independent groups

Distribution	Parameter	Statistical test
Normal	Mean	- Student t-test with equal variance - Student t-test with unequal variance
Non-normal	Median	- Mann-Whitney test, - Quantile regression

Example I:

- ➲ Researchers wanted to test if means/median of **weights** of HIV patients who received NVP, and HIV patients who received EFV, are different.

Variance ratio test

```
. sdtest bw,by(group)
Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     NVP |      70    54.82286    1.034775    8.657545    52.75854    56.88718
     EFV |      70    54.36429    1.266647    10.59753    51.83739    56.89118
---------+--------------------------------------------------------------------
combined |     140    54.59357    .8150796    9.644152    52.98201    56.20513
------------------------------------------------------------------------------
    ratio = sd(NVP) / sd(EFV)                                     f =   0.6674
Ho: ratio = 1                                    degrees of freedom =   69, 69
    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.0477         2*Pr(F < f) = 0.0954           Pr(F > f) = 0.9523
```

Conclusion

➲ Variances between two groups are not different.

Student t-test with equal variance

```
. ttest bw,by(group)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     NVP |      70    54.82286    1.034775    8.657545    52.75854    56.88718
     EFV |      70    54.36429    1.266647    10.59753    51.83739    56.89118
---------+--------------------------------------------------------------------
combined |     140    54.59357    .8150796    9.644152    52.98201    56.20513
---------+--------------------------------------------------------------------
    diff |            .4585714    1.635589               -2.775485    3.692628
------------------------------------------------------------------------------
    diff = mean(NVP) - mean(EFV)                                  t =   0.2804
Ho: diff = 0                                     degrees of freedom =      138
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.6102         Pr(|T| > |t|) = 0.7796          Pr(T > t) = 0.3898
```

Conclusion ➲ Mean weights between two groups are not different.

Example II:

➲ Researchers wanted to test if CD4 count of HIV patients who received NVP, and HIV patients who received EFV, are different.

Quantile regression

```
. xi:qreg cd4c i.group
i.group           _Igroup_1-2         (naturally coded; _Igroup_1 omitted)
Iteration  1:  WLS sum of weighted deviations =  7527.2521

Iteration  1: sum of abs. weighted deviations =       7546
Iteration  2: sum of abs. weighted deviations =       7178
Iteration  3: sum of abs. weighted deviations =       6784

Median regression                                    Number of obs =        140
  Raw sum of deviations     6802 (about 29)
  Min sum of deviations     6784                     Pseudo R2     =     0.0026

------------------------------------------------------------------------------
        cd4c |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Igroup_2 |         -7   11.96985    -0.58   0.560    -30.66803    16.66803
       _cons |         36   8.463962     4.25   0.000     19.26418    52.73582
------------------------------------------------------------------------------
```

Conclusion

➲ **Median of CD4 count between two groups are not different.**

Dummy table for two groups comparison

Characteristics	NVP	EFV	P value
	Mean (SD)	Mean (SD)	
Age (year)			
Weight (kg)			
Height (cm)			
BMI (kg/m^3)			
CD4 count; median (range)			

Paired sample

- Comparison of systolic blood pressure before and after used of OC in pre-menopausal women.
- In matched case-control study, matched by age and sex, which aim to compare oral hygiene index between periodontitis and non-periodontitis patients.

Statistical test for paired sample

Distribution	Parameter	Statistical test
Normal	Mean	Paired t-test
Non-normal	Median	Wilcoxon matched signed-rank test

Example III:

➲ Researchers wanted to test if mean weights of HIV patients before and after receiving an antiretroviral therapy regimen are different.

Paired t-test

```
. ttest bw0= bw12
Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     bw0 |     121    54.56694    .8926941    9.819635    52.79947    56.33441
    bw12 |     121    57.31322    .9380435    10.31848    55.45596    59.17048
---------+--------------------------------------------------------------------
    diff |     121   -2.746281    .3710625    4.081688   -3.480959   -2.011603
------------------------------------------------------------------------------
     mean(diff) = mean(bw0 - bw12)                                t =  -7.4011
 Ho: mean(diff) = 0                              degrees of freedom =      120
 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Conclusion

➲ **Mean weights before and after receiving regimen are different.**

Example IV:

- Researchers wanted to test if median of CD4 count of HIV patients before and after receiving an antiretroviral therapy regimen are different.

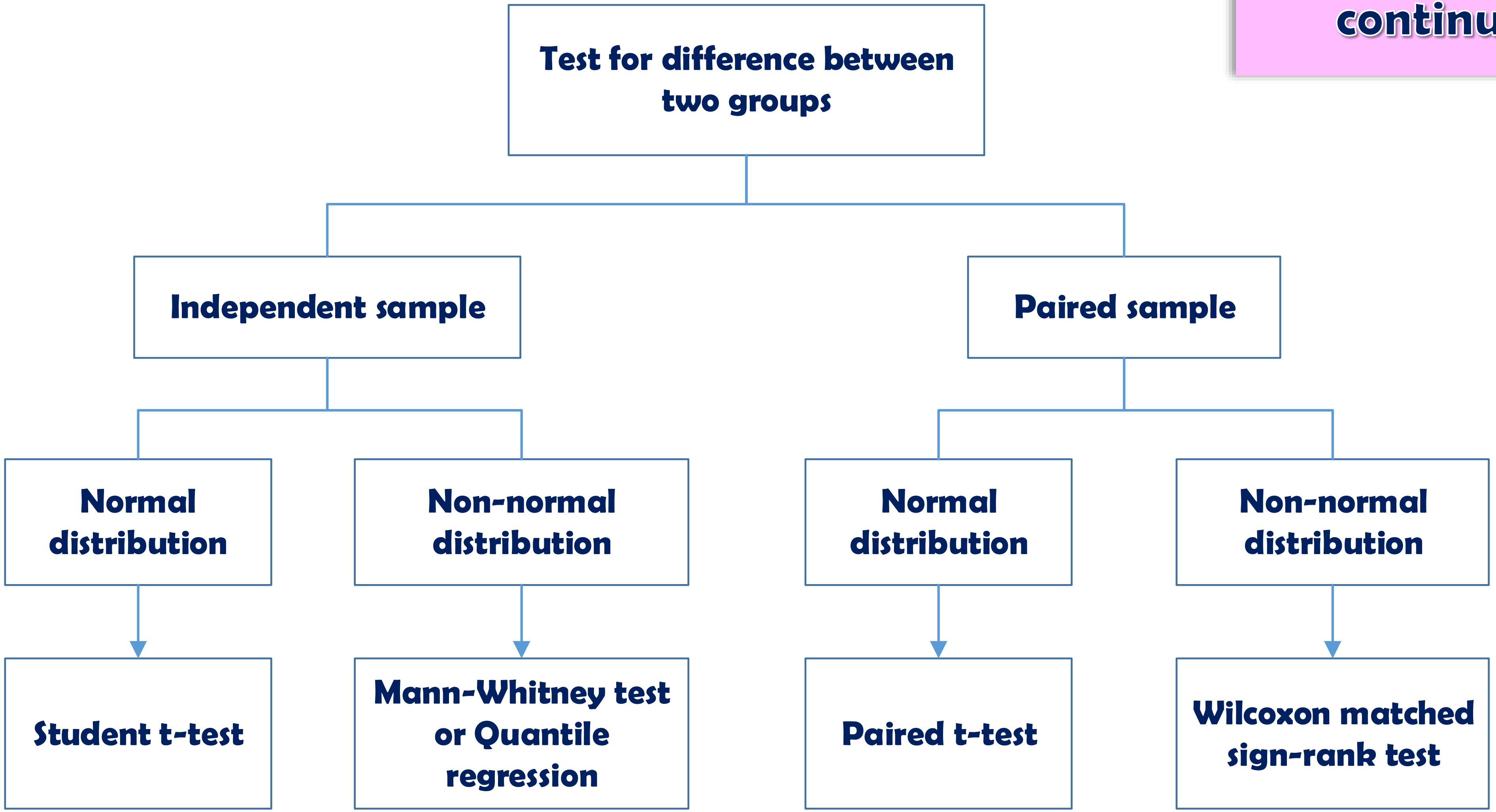
Wilcoxon matched signed-rank test

```
. signrank  cd4c0= cd4c12
          Wilcoxon signed-rank test
                 sign |      obs   sum ranks    expected
        -------------+---------------------------------
             positive |        7       256.5        3570
             negative |      112      6883.5        3570
                 zero |        0           0           0
        -------------+---------------------------------
                  all |      119        7140        7140
        unadjusted variance   142205.00
        adjustment for ties       -5.38
        adjustment for zeros       0.00
                             ----------
        adjusted variance     142199.63
        Ho: cd4c0 = cd4c12
                     z =  -8.787
            Prob > |z| =   0.0000
```

Conclusion

- Reject null hypothesis
- Median CD4 count before and after receiving regimen are different

Hypothesis testing for continuous data



Test for difference between two groups

Independent sample

Normal distribution

Student t-test

Non-normal distribution

Mann-Whitney test or Quantile regression

Paired sample

Normal distribution

Paired t-test

Non-normal distribution

Wilcoxon matched sign-rank test

Thank You
for your attention